ตัวแบบสำหรับการเรียงลำดับผลลัพธ์การค้นคืนในระบบค้นคืนบทความวิจัย โดยการใช้ข้อมูลทางบรรณานุกรม

ขวัญเรือน โสอุบล

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรวิทยาศาสตรมหาบัณฑิต
สาขาวิชาวิศวกรรมเว็บ คณะเทคโนโลยีสารสนเทศ
มหาวิทยาลัยธุรกิจบัณฑิตย์
พ.ศ. 2557

A Model for Ranking Search Results in a Research Paper Search Engine Using Bibliographic Information

Khwanruan So-Ubol

A Thesis Submitted in Partial Fulfillment of the Requirements

for the Degree of Master of Science in Web Engineering

Faculty of Information Technology, Dhurakij Pundit University

2014

ใบรับรองวิทยานิพนธ์

คณะเทคโนโลยีสารสนเทศ มหาวิทยาลัยธุรกิจบัณฑิตย์

ปริญญา วิทยาศาสตรมหาบัณฑิต

หัวข้อวิทยานิพนธ์ ตัวแบบสำหรับการเรียงลำดับผลลัพธ์การค้นคืนในระบบค้นคืนบทความวิจัยโดยการใช้ข้อมูลทางบรรณานุกรม

เสนอโดย นางสาวขวัญเรือน โสอุบล

สาขาวิชา วิศวกรรมเว็บ

อาจารย์ที่ปรึกษาวิทยานิพนธ์ ผู้ช่วยศาสตราจารย์ ดร.วรสิทธิ์ ชูชัยวัฒนา

ได้พิจารณาเห็นชอบโดยคณะกรรมการสอบวิทยานิพนธ์แล้ว

.....ประธานกรรมการ

(รองศาสตราจารย์ ดร.นุชรี เปรมชัยสวัสดิ์)

.....กรรมการและอาจารย์ที่ปรึกษาวิทยานิพนธ์

(ผู้ช่วยศาสตราจารย์ ดร.วรสิทธิ์ ชูชัยวัฒนา)

..... กรรมการ

(อาจารย์ ดร.นันทิกา ปริญญาพล)

..... กรรมการ

(อาจารย์ ดร.พิจิตรา จอมศรี)

..... กรรมการ

(อาจารย์ ดร.ศิริลักษณ์ อารีรัชชกุล)

คณะเทคโนโลยีสารสนเทศรับรองแล้ว

..... คณบดีคณะเทคโนโลยีสารสนเทศ

(รองศาสตราจารย์ ดร.นุชรี เปรมชัยสวัสดิ์)

วันที่ 8 เดือน สิงหาคม พ.ศ. 2557

หัวข้อวิทยานิพนธ์	ตัวแบบสำหรับการเรียงลำดับผลลัพธ์การค้นคืนในระบบค้นคืนบทความวิจัยโดยการใช้ข้อมูลทางบรรณานุกรม
ชื่อผู้เขียน	ขวัญเรือน โสอุบล
อาจารย์ที่ปรึกษา	ผู้ช่วยศาสตราจารย์ ดร.วรสิทธิ์ ชูชัยวัฒนา
สาขาวิชา	วิศวกรรมเว็บ
ปีการศึกษา	2556

บทคัดย่อ

อินเทอร์เน็ตและเวิลด์ไวด์เว็บเป็นช่องทางใหม่สำหรับการจัดเก็บและเผยแพร่สารสนเทศ นักวิจัยจากทั่วโลกมักจะทำการค้นหาบทความวิชาการและบทความวิจัยที่น่าสนใจ ผ่านระบบห้องสมุดดิจิทัลออนไลน์ เช่น IEEE Explore และ ACM Digital Library อย่างไรก็ตามการเรียงลำดับผลลัพธ์การค้นคืนของระบบดังกล่าวจะเป็นพิจารณาจากการเปรียบเทียบความเหมือนระหว่างคำสืบค้นและดัชนีของเอกสาร ซึ่งเทคนิคนี้เรียกว่า Query Dependent Ranking

งานวิจัยนี้มีวัตถุประสงค์เพื่อนำเสนอตัวแบบสำหรับการเรียงลำดับผลลัพธ์การค้นคืนบทความวิจัยโดยการใช้ข้อมูลทางบรรณานุกรม ในตัวแบบดังกล่าว นอกจากจะพิจารณาความเหมือนระหว่างคำสืบค้นและดัชนีของเอกสารแล้ว ยังพิจารณาปัจจัยที่เกี่ยวข้องกับคุณภาพของบทความวิจัยและแหล่งตีพิมพ์ด้วย คุณภาพของบทความวิจัยพิจารณาจากจำนวนของบรรณานุกรม จำนวนของการถูกอ้างอิงโดยบทความอื่น และคุณภาพของแหล่งตีพิมพ์

การทดลองได้ถูกจัดขึ้น โดยมีการใช้ NDCG และ MAP เป็นตัววัดสำหรับการประเมินประสิทธิผลของตัวแบบการเรียงลำดับผลลัพธ์การค้นคืนที่เสนอในงานวิจัยชิ้นนี้ ผลการทดลองพบว่าผลการเรียงลำดับของการค้นคืนที่ได้จากตัวแบบมีประสิทธิผลมากกว่า ดังนั้นคุณภาพของบทความวิจัยและคุณภาพของแหล่งตีพิมพ์มีส่วนช่วยในงานการเรียงผลลัพธ์การค้นคืนของบทความวิจัย

Thesis Title	A Model for Ranking Search Results in a Research Paper Search Engine Using Bibliographic Information
Author	Khwanruan So-Ubol
Thesis Advisor	Assistant Professor Dr.Worasit Choochaiwattana
Academic Program	Web Engineering
Academic Year	2013

ABSTRACT

The Internet and World Wide Web provide people a new way to store and disseminate information. Researchers from all over the world always search for interesting academic papers via online digital libraries such as IEEE Explore and ACM Digital library. However, a ranking of search results from these systems determines by comparing matches between query terms and document indexes. This technique is called *Query Dependent Ranking*.

This research aims at proposing a model for ranking research paper search results using bibliographic information. Instead of determining only the matches between query terms and documents indexes, the proposed ranking model also considers a quality feature such as a quality of research papers and a quality of publishers. The quality of the papers determines by a number of reference and a number of citations and the quality of the publishers.

The experiment was conducted. NDCG and MAP were used as a metric to evaluate the effectiveness of the proposed ranking model. The result showed that the proposed ranking model provide better search results ranking. Thus, the quality of research papers and the quality of publishers contribute to research paper ranking tasks.

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้สำเร็จลุล่วงไปได้ด้วยความอนุเคราะห์และการสนับสนุนตลอดการดำเนินการวิจัยจาก ผศ.ดร.วรสิทธิ์ ชูชัยวัฒนา อาจารย์ที่ปรึกษาวิทยานิพนธ์ ที่กรุณาให้คำแนะนำความรู้ ความคิดเห็นต่างๆ อันเป็นประโยชน์ในการทำวิจัย

ขอกราบขอบพระคุณคณาจารย์สาขาวิศวกรรมเว็บ คณะเทคโนโลยีสารสนเทศมหาวิทยาลัยธุรกิจบัณฑิตย์ ที่กรุณาถ่ายทอดความรู้อันเป็นประโยชน์ตลอดการศึกษา

ขอขอบคุณเพื่อนๆ ทุกคนที่คอยให้ความช่วยเหลือ เอื้อเฟื้อด้านต่างๆ รวมถึงกำลังใจที่คอยแบ่งปันให้กันตลอดเวลา

สุดท้ายขอขอบคุณกำลังใจจากครอบครัว ซึ่งเป็นพลังอันสำคัญและยิ่งใหญ่ที่คอยผลักดันให้การทำวิทยานิพนธ์ครั้งนี้สำเร็จลุล่วงไปได้ด้วยดี

ขวัญเรือน โสอุบล

สารบัญ

	หน้า
บทคัดย่อภาษาไทย	ฆ
บทคัดย่อภาษาอังกฤษ	ง
กิตติกรรมประกาศ	จ
สารบัญตาราง	ช
สารบัญภาพ	ฌ
บทที่	
1. บทนำ	1
1.1 ที่มาและความสำคัญของงาน	1
1.2 วัตถุประสงค์ของการศึกษา	2
1.3 สมมติฐานของการวิจัย	2
1.4 ประโยชน์ที่คาดว่าจะได้รับ	2
1.5 ขอบเขตของการวิจัย	2
1.6 นิยามคำศัพท์	3
2. ทฤษฎีและงานวิจัยที่เกี่ยวข้อง	5
2.1 ทฤษฎี	5
2.2 งานวิจัยที่เกี่ยวข้อง	10
3. ระเบียบวิจัย	15
3.1 ทฤษฎีการวิเคราะห์ปัญหาและศึกษาค้นคว้าข้อมูล	15
3.2 เครื่องมือที่ใช้ในการวิจัย	24
4. ผลการดำเนินงาน	25
4.1 ค่าเฉลี่ย NDCG	25
4.2 ค่าเฉลี่ย MAP	26
5. สรุป อภิปรายผล และข้อเสนอแนะ	28
5.1 สรุปและอภิปรายผล	28
5.2 ปัญหาและอุปสรรค	29
5.3 ข้อเสนอแนะ	29

สารบัญ (ต่อ)

	หน้า
บรรณานุกรม.....	31
ภาคผนวก	
ก ตัวอย่างการเตรียมคลังเอกสาร.....	36
ข การออกแบบตารางฐานข้อมูล.....	45
ค ตัวอย่างหน้าจอระบบค้นคืนบทความวิจัย.....	48
ง ตัวอย่างผลการประเมินจากผู้ทดสอบ.....	50
จ บทความการประชุมวิชาการระดับชาติ ด้านคอมพิวเตอร์และเทคโนโลยีสารสนเทศ (NCCIT) ครั้งที่10.....	53
ประวัติผู้เขียน.....	60

สารบัญตาราง

ตารางที่ **หน้า**

2.1 ตัวอย่างบทความวิจัย..... 5
2.2 ขั้นตอนการประมวลผลเอกสาร..... 6
3.1 คลังเอกสาร..... 16
3.2 ฟิลด์ข้อมูลที่ใช้ทำดัชนี..... 18
3.3 คุณลักษณะของตัวแบบ..... 21
3.4 Judgments Score 23
4.1 ค่าเฉลี่ย NDCG 26
4.2 ค่าเฉลี่ย MAP..... 27

สารบัญภาพ

ภาพที่ หน้า

2.1 ปริภูมิเว็กเตอร์เอกสาร..... 7
2.2 Term –document matrix ของเอกสาร..... 8
2.3 เว็กเตอร์ของเอกสารและคำค้น..... 8
3.1 ขั้นตอนการทำงานของระบบค้นคืนเอกสาร..... 15
3.2 ขั้นตอนการวิเคราะห์คำเพื่อสร้างและค้นคืนผ่านดัชนี..... 17
3.3 กรอบแนวคิดการสร้างตัวแบบ..... 20
3.4 การคำนวณ Hybrid Score 22
3.5 ขั้นตอนการประเมินผล..... 24
4.1 ค่าเฉลี่ย NDCG 26
4.2 ค่าเฉลี่ย MAP..... 27

บทที่ 1
บทนำ

1.1 ที่มาและความสำคัญของงาน

เทคโนโลยีสารสนเทศและอินเทอร์เน็ตถูกพัฒนาไปอย่างรวดเร็ว ทำให้ปริมาณข้อมูลและสารสนเทศต่างๆ ถูกเผยแพร่มากมายมหาศาล ดังนั้นการพัฒนาระบบค้นคืนข้อมูลที่มีประสิทธิภาพ และตรงกับความต้องการของผู้ใช้จึงทำได้ยากและมีความจำเป็นมากขึ้น ระบบค้นคืนที่นำมาใช้งานในอดีต เช่น การสืบค้นข้อมูลของ Yahoo! ใช้วิธีการที่เรียกว่า Catalog Based Information Retrieval จะเป็นการสืบค้นจากหมวดหมู่หลักแล้วค่อยๆ ย่อยลงไปจนถึงหัวข้อที่ต้องการ หรืออีกวิธีการหนึ่งที่นิยมใช้กันอย่างแพร่หลายก็คือ Query Based Search Engine เป็นการสืบค้นข้อมูลที่มักจะพิจารณา โดยเปรียบเทียบความเหมือนระหว่างคำค้น (Query) กับคำที่ปรากฏอยู่ในเอกสารเท่านั้น ซึ่งผลลัพธ์การค้นคืนที่ได้ส่วนใหญ่มักจะไม่มีความสัมพันธ์กับคำค้นที่ต้องการ วิธีการดังกล่าวถูกเรียกว่า Query Dependent Ranking หรือ Similarity Ranking สำหรับระบบค้นคืนในยุคถัดมาคือ กูเกิ้ล (Google) เริ่มมีการนำเอาปัจจัยที่เกี่ยวข้องกับเอกสารอื่นๆ มาพิจารณาร่วมด้วย ตัวอย่างเช่น คุณภาพของเอกสาร การเชื่อมโยงระหว่างเอกสารที่อยู่ในเครือข่าย เป็นต้น พบว่าให้ผลลัพธ์การค้นคืนและการเรียงลำดับที่น่าพึงพอใจกับผู้ใช้มากขึ้น เมื่อดูจากสถิติการเข้าใช้งานมากที่สุดถึง 68% โดยวิธีการดังกล่าวถูกเรียกว่า Query Independent Ranking หรือ Static Ranking

อีกปัญหาหนึ่งที่มักจะขึ้นกับนักวิจัยคือการสืบค้นหรือการค้นหาบทความวิจัยที่ไม่ตรงกับต้องการ เนื่องจากปริมาณบทความวิจัยที่มีอยู่ในระบบ ไม่ว่าจะเป็นระบบของ IEEE Explore และ ACM Digital Library ยังใช้การค้นคืนบทความแบบ Query Dependent Ranking เปรียบเทียบคำเหมือนระหว่างคำค้นกับฐานข้อมูลของระบบนั่นเอง

ในการศึกษางานวิจัยนี้ทำการทดลองเพื่อพิสูจน์สันนิษฐานว่า เทคนิคสำหรับการเรียงลำดับแบบ Query Independent Ranking แบบผสมผสานข้อมูลบรรณานุกรม ให้ผลลัพธ์ดีกว่า Query Dependent Ranking เพียงอย่างเดียว โดยการสร้างดัชนีต้นแบบของทั้งสองวิธี และเพื่อให้ผลการทดลองสามารถควบคุมปัจจัยที่มีผลกระทบต่อการวิจัยและขอบเขตของข้อมูล โดยมุ่งเน้น

ไปที่ข้อมูลบทความวิจัยทางด้านวิทยาการคอมพิวเตอร์ ซึ่งจะกล่าวถึงขอบเขต และกลุ่มประชากรในหัวข้อถัดไป

1.2 วัตถุประสงค์ของการศึกษา

สร้างตัวแบบสำหรับการเรียงลำดับผลลัพธ์การค้นคืนบทความวิจัยบนระบบค้นคืน

1.3 สมมติฐานของการวิจัย

จากที่มาของปัญหาของระบบค้นคืนบทความวิจัยที่กล่าวไปข้างต้น จึงมีแนวความคิดว่า ถ้ามีการนำข้อมูลบรรณานุกรมของบทความวิจัยเอกสารส่วนที่เป็นคุณสมบัติของบทความวิจัย (Paper) เรียกว่าเป็นการวัดคุณภาพของบทความ และคุณภาพของของแหล่งตีพิมพ์ (Publisher) เข้ามาพิจารณาร่วมกับการทำ Query Dependent Ranking จะให้ผลลัพธ์การค้นคืนที่ดีและการเรียงลำดับที่ดีขึ้น

1.4 ประโยชน์ที่คาดว่าจะได้รับ

สามารถนำตัวแบบที่ได้จากการวิจัยมาปรับปรุงคุณภาพในระบบค้นคืนบทความวิจัย ที่มีคุณภาพ ซึ่งจะส่งผลให้ผู้ใช้มีความพึงพอใจเพิ่มมากขึ้นด้วย

1.5 ขอบเขตของการวิจัย

1.5.1 ขอบเขตด้านเนื้อหา

1.5.1.1 เก็บรวบรวมข้อมูลบทความวิจัยจากเว็บ http://academic.research.microsoft.com เป็นผู้ให้บริการข้อมูลเกี่ยวกับบทความวิจัยด้านต่างๆ บรรณานุกรม และแหล่งตีพิมพ์

1.5.1.2 ครอบคลุมงานวิจัยด้านกระบวนการค้นคืนเอกสาร (Search Engine) เว็บแบบปรับเหมาะ (Adaptive Web) ปัญญาประดิษฐ์ (Artificial Intelligent) การทำเหมืองข้อมูล (Data Mining) วิศวกรรมซอฟต์แวร์ (Software Engineering) การปฏิสัมพันธ์ระหว่างคอมพิวเตอร์กับมนุษย์ (Human Computer Interaction: HCI) โครงข่ายประสาทเทียม (Neural Network)

1.5.2 ขอบเขตด้านประชากร

ประชากรตัวอย่างเป็นนักศึกษาปริญญาโท ปริญญาเอก อาจารย์ และนักวิจัยที่ศึกษาและทำงานวิจัยที่อยู่ในขอบเขตของเนื้อหาในหัวข้อ 1.5.1.2 เท่านั้น เพื่อให้กลุ่มประชากรสามารถเข้าใจเนื้อหาบทความวิจัยและสามารถประเมินผลได้อย่างถูกต้อง

1.5.3 ขอบเขตด้านเวลา

ช่วงเวลาที่เก็บข้อมูลคือเดือนมิถุนายน - สิงหาคม พ.ศ. 2556

1.6 นิยามคำศัพท์

เสิร์ชเอนจิน (Search Engine) หมายถึง ระบบค้นคืนหรือโปรแกรมที่ช่วยในการสืบค้นข้อมูล โดยเฉพาะข้อมูลบนอินเทอร์เน็ต โดยครอบคลุมทั้งข้อความ รูปภาพ ภาพเคลื่อนไหว เพลง ซอฟต์แวร์ แผนที่ ข้อมูลบุคคล กลุ่มข่าว และอื่นๆ ซึ่งแตกต่างกันไปแล้วแต่โปรแกรมหรือผู้ให้บริการแต่ละราย เสิร์ชเอนจินส่วนใหญ่จะค้นหาข้อมูลจากคำสำคัญ (Keywords) ที่ผู้ใช้ป้อนเข้าไป จากนั้นก็จะแสดงรายการผลลัพธ์การค้นคืนเอกสาร

ครอเลอร์ (Crawler) หมายถึง โปรแกรมหรือแอปพลิเคชันที่ทำหน้าที่สแกนและอ่านข้อมูลจากเว็บไซต์ โดยเข้าถึงจากลิงค์ของเว็บหนึ่งไปยังเว็บอื่นๆ เพื่อนำข้อมูลที่ได้มาสร้างดัชนีสำหรับระบบค้นคืน

คลังเอกสาร (Document Corpus) หมายถึง ฐานข้อมูลหรือที่เก็บรวมรวบเอกสารของระบบค้นคืน

การวิเคราะห์คำ (Parsing) หมายถึง การวิเคราะห์เอกสาร HTML ที่ได้จากการ Crawl ตามโครงสร้าง เพื่อสกัดข้อมูลที่ต้องการให้อยู่ในรูปแบบของฟิลด์ (Field)

โทเคนไนซิ่ง (Tokenizing) หมายถึง กระบวนการประมวลผลข้อความในเอกสารเพื่อให้อยู่ในรูปแบบของคำ ซึ่งในการวิจัยนี้จะหมายถึงอักขระที่มีความยาวตั้งแต่ 3 ตัวขึ้นไป ตัดคำด้วยเว้นวรรค (space) หรืออักขระพิเศษ และอักขระทั้งหมดถูกแปลงเป็นอักขระตัวเล็ก

สต็อปเวิร์ด (Stop words) หมายถึง คำที่ใส่เพิ่มเติมทำหน้าที่ขยายหรือเป็นส่วนประกอบของคำอื่นๆ ตัวอย่าง article ในภาษาอังกฤษ เช่น the, a, an คำที่บอกถึงปริมาณ เช่น over, under, above, below เป็นต้น คำเหล่านี้เป็นคำที่ปรากฏอยู่ในทุกๆ เอกสาร และไม่สามารถนำมาใช้บ่งบอกถึงความเกี่ยวข้องระหว่างเอกสารกับคำค้นได้

สเตมมิ่ง (Stemming) หมายถึง กระบวนการประมวลผลคำหนึ่งที่อยู่ในรูปแบบต่างๆ ตามหน้าที่ของคำ เช่น คำคุณศัพท์ คำนาม กริยา เป็นต้น ลดรูปให้กลายเป็นรากศัพท์ของคำ

ลูซีน (Lucene) หมายถึง ไลบรารี่สำหรับวิเคราะห์เอกสารให้อยู่ในรูปของดัชนีเพื่อให้การสืบค้นทำได้รวดเร็วมากยิ่งขึ้น โดยผ่านกระบวนการทำ Tokenizing, Stop words และ Stemming

ดัชนี (Index) หมายถึง โครงสร้างข้อมูลที่ใช้แทนคำ หรือ ตัวเลข ในเอกสารในรูปแบบของสัญลักษณ์ (Signature) ไปยังตำแหน่งของเนื้อหาที่อยู่ในเอกสาร เซตของเอกสาร หรือแฟ้มฐานข้อมูล โดยใช้ตัวชี้ (Pointer) ภายในดัชนีจะมีการเรียงลำดับเนื้อหา เลขหน้า ตัวชี้ และข้อมูลอื่นๆ เพื่อให้สามารถเข้าถึงหรือค้นหาข้อความในเอกสารได้เร็วขึ้น

ค่าน้ำหนักของคำ (Term Weight) หมายถึง ค่าน้ำหนักที่บ่งบอกถึงความสำคัญของคำ แต่ละคำที่อยู่ในคลังเอกสาร จะถูกปรับค่าตามอัตราส่วนระหว่างจำนวนเอกสารทั้งหมดกับจำนวนเอกสารที่มีคำนี้ปรากฏอยู่

บทที่ 2
ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

ในบทนี้จะกล่าวถึงทฤษฎี แนวคิด องค์ความรู้ต่างๆ ที่เกี่ยวข้องรวมถึงงานวิจัยที่ผ่านมาสำหรับการดำเนินการวิจัยและประเมินผล ซึ่งแบ่งออกเป็น 2 ส่วนคือทฤษฎีและงานวิจัยที่เกี่ยวข้องกับระบบค้นคืนในด้านต่างๆ มีการเทคนิคและแนวความคิดเกี่ยวกับการเรียงลำดับ มีรายละเอียดดังนี้

2.1 ทฤษฎี

2.1.1 ตัวแบบการค้นคืนแบบบูลีน (Boolean Retrieval Model)

เป็นวิธีการค้นคืนเอกสารที่ถูกนำมาใช้ในงานด้าน Information Retrieval หลักการสำคัญคือ การเปรียบเทียบคำค้นกับเอกสารแบบ Exact-match Retrieval หรือจะเรียกว่าเป็นการนำคำค้นมาเปรียบเทียบกับเอกสารทีละอักขระ โดยเอกสารที่เป็นผลลัพธ์จะต้องมีคำค้นปรากฏอยู่ในเอกสารเสมอ ค่าที่วัดออกมาได้จะอยู่ในรูปแบบไบนารี (Binary) คือ จริง (TRUE) หรือเท็จ (FALSE) เท่านั้น โดยวิธีการเปรียบเทียบระหว่างเอกสารกับคำค้นมีโอเปอเรเตอร์ 3 แบบ ได้แก่ AND OR และ NOT

ตารางที่ 2.1 ตัวอย่างบทความวิจัย

เอกสาร	ฟิลด์	เนื้อหา
1	Title	**Adaptive** web caching: towards a new global caching architecture
	Abstract	An **adaptive**, highly scalable, and robust web caching system is needed to effectively handle the exponential growth and extreme dynamic environment of the World Wide Web.
	Keywords	Exponential Growth, Self Organization, Smooth Transition, Web Caching, and Forward, Local Group, World Wide Web

ตารางที่ 2.1 (ต่อ)

เอกสาร	ฟิลด์	เนื้อหา
2	Title	**Adaptive** Web Sites: an AI Challenge
	Abstract	We challenge the AI community to create **adaptive** web sites: sites that automatically improve their organization and presentation based on user access data.
	Keywords	**Adaptive** Web Site, Machine Learning, Plan Recognition, User Interaction, User Interface Design, User Model
3	Title	Relational Markov models and their application to **adaptive** web navigation
	Abstract	Relational Markov models (RMMs) are a generalization of Markov models where states can be of different types.
	Keywords	Crossed Product, Web Mining, Web Navigation**, Adaptive** Web

จากตัวอย่างบทความวิจัยตามตารางที่ 2.1 ในคลังเอกสารมี 3 เอกสาร การค้นคืนกำหนดให้เงื่อนไขในการค้นคืนเอกสาร คือ Title AND Keyword OR Abstract คำค้น คือ "Adaptive" จะได้ผลลัพธ์การค้นคืนตามตารางที่ 2.2 คือเอกสาร 2 และ 3 เท่านั้น เนื่องจากในเอกสาร 1 ไม่มีคำว่า "Adaptive" ปรากฏอยู่ใน Keyword แต่เมื่อพิจารณาเข้าไปในเนื้อหาของเอกสาร 1 จะพบว่าเนื้อหาในเอกสารก็มีความเกี่ยวข้องและสัมพันธ์กับคำค้นด้วยเหมือนกัน ซึ่งจะเห็นว่าวิธีการนี้จะต้องสแกนคำที่อยู่ในเอกสารทุกเอกสาร ทำให้ขาดทั้งประสิทธิภาพและประสิทธิผลในการสร้างระบบค้นคืนที่มีปริมาณเอกสารจำนวนมากและให้การตอบสนองที่เร็ว

ตารางที่ 2.2 ขั้นตอนการประมวลผลเอกสาร

เอกสาร	Title	Keywords	Abstract	ผลลัพธ์
1	T	F	T	F
2	T	T	T	T
3	T	T	F	T

2.1.2 ตัวแบบการค้นคืนแบบปริภูมิเว็กเตอร์ (Vector Space Model)

แนวความคิดของเว็กเตอร์คือ การใช้เว็กเตอร์แต่ละมิติ (Dimension) เป็นตัวแทนของเอกสารและคำค้น จากสมการที่ 2.1 แทนเอกสาร

$$D_i = (d_{i1}, d_{i2}, \dots, d_{it},) \tag{2.1}$$

เมื่อ

D_i แทนเว็กเตอร์ของ Index Term

t แทนจำนวน Index Term เช่น คำ (Words) สเตม (Stems) วลี (Phrases) และอื่นๆ

d_{ij} แทนค่าน้ำหนักของ Term ที่ตำแหน่ง j

เมื่อคลังเอกสารมีจำนวน n เอกสาร เขียนปริภูมิเว็กเตอร์ด้วยเมตริกค่าน้ำหนักของคำตามภาพที่ 2.1 ได้ดังนี้

	$Term_1$	$Term_2$	...	$Term_1$
Doc_1	d_{11}	d_{12}	...	d_{1t}
Doc_2	d_{21}	d_{22}	...	d_{2t}
⋮	⋮			
Doc_n	d_{n1}	d_{n2}	...	d_{nt}

ภาพที่ 2.1 ปริภูมิเว็กเตอร์เอกสาร

ในลักษณะเดียวกัน แทนคำค้น Q ด้วยเว็กเตอร์ของ Term Weight เขียนเซตของคำในเอกสาร หรือคำในคำค้นได้ตามสมการที่ 2.2

$$Q = (q_1, q_2, \dots, q_t,) \tag{2.2}$$

เมื่อ

Q แทนคำค้น

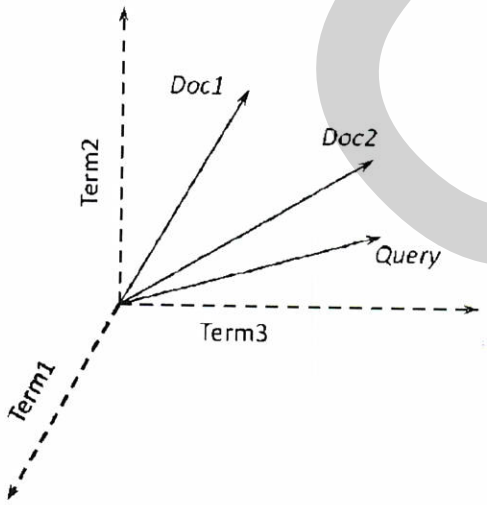
t แทนค่าน้ำหนักของคำในเอกสารหรือในคำค้น

จากภาพที่ 2.2 เป็นตัวอย่างของปริภูมิเว็กเตอร์ของเอกสารกับจำนวน Term ที่ปรากฏอยู่ในแต่ละเอกสาร เมื่อแต่ละแถวคือค่าน้ำหนักของคำ (Term) และแต่ละคอลัมน์คือเอกสาร เมื่อนำเว็กเตอร์มาใช้แทนเอกสารและคำ สามารถหมุนแกนทั้งสองได้ตามความเหมาะสม ตัวอย่างเอกสาร D_3 แทนด้วยเว็กเตอร์ (1, 1, 0, 2, 0, 1, 0, 1, 0, 0, 1) และคำค้น "Tropical Fish" แทนด้วยเว็กเตอร์ Q (0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1) เพื่อให้เข้าใจง่ายขึ้นแทนเว็กเตอร์ด้วยภาพ 3 มิติได้ตามภาพที่ 2.3

D_1 Tropical Freshwater Aquarium Fish.
D_2 Tropical Fish, Aquarium Care, Tank Setup.
D_3 Keeping Tropical Fish and Goldfish in Aquariums, and Fish Bowls.
D_4 The Tropical Tank Homepage - Tropical Fish and Aquariums.

Terms	Documents			
	D_1	D_2	D_3	D_4
aquarium	1	1	1	1
bowl	0	0	1	0
care	0	1	0	0
fish	1	1	2	1
freshwater	1	0	0	0
goldfish	0	0	1	0
homepage	0	0	0	1
keep	0	0	1	0
setup	0	1	0	0
tank	0	1	0	1
tropical	1	1	1	2

ภาพที่ 2.2 Term –document matrix ของเอกสาร



ภาพที่ 2.3 เว็กเตอร์ของเอกสารและคำค้น

ซึ่งในความเป็นจริงแล้วมิติของทั้งเอกสารและ Term เองนั้น มีปริมาณมหาศาล ทวีคูณมากกว่าจำนวนเอกสารเกินกว่าที่จะแสดงออกมาเป็นภาพสามมิติได้ เพื่อให้ง่ายต่อการคำนวณ จึงต้องแทนแต่ละมิติด้วยจุด (Point) แล้ววัดระยะห่าง (Distance) ระหว่างมุมของเว็กเตอร์ คิดได้จากสมการที่ 2.3 เรียกค่านี้ว่า Similarity Measure หรือ Cosine Similarity Ranking เป็นผลรวมของ Dot Product ระหว่าง Term Weight ของเอกสารกับคำค้น โดยทำ Normalized ค่าคะแนนนี้ด้วย Product Length ของเว็กเตอร์ทั้งสอง นั่นหมายความว่าถ้าระยะห่างมีค่าเข้าใกล้ศูนย์ หรือเป็นศูนย์ แสดงว่าคำที่อยู่ในเอกสารสองเอกสารหรือคำค้นไม่มีความเกี่ยวข้องระหว่างกัน

$$Cosine(D_i, Q) = \frac{\sum_{j=1}^{t} d_{ij} . q_j}{\sqrt{\sum_{j=1}^{t} {d_{ij}}^2 . \sum_{j=1}^{t} {q_j}^2}} \tag{2.3}$$

ปัจจัยที่มีผลกับ Vector Space Model คือ Term ที่ปรากฏอยู่ในเอกสารและจำนวน Term ที่ตรงกับคำค้น ซึ่งจะเห็นว่าการที่เอกสารมีความยาวที่มากกว่าย่อมมีจำนวน Term ที่มากกว่า เพื่อลดผลกระทบที่เกิดขึ้น จำเป็นต้องนำความยาวเอกสารมาพิจารณาเพิ่มคือ Term Frequency และจำนวนเอกสารที่ Term นั้นปรากฏ เรียกว่าค่า Term Weights ค่าน้ำหนักของคำนี้คิดจาก $tf_{ik} . idf_i$ พิจารณาเป็น 2 ค่าด้วยกันคือ

Term Frequency (*tf*) ค่าความถี่ของคำในเอกสาร บ่งบอกถึงความสำคัญของคำที่อยู่ในเอกสารนั้น คำนวณได้จากสมการที่ 2.4

$$tf_{ik} = \frac{f_{ik}}{\sum_{j=1}^{t} f_{ij}} \tag{2.4}$$

เมื่อ

tf_{ik} แทน Term Frequency Weight ของคำ k ในเอกสาร D_i

f_{ik} แทน จำนวนครั้งที่คำ k ปรากฏในเอกสาร D_i

Inverse document frequency (*idf*) เป็นการพิจารณาถึงความสำคัญของคำที่อยู่ในคลังเอกสาร โดยดูจาก Term นั้นปรากฏอยู่ในเอกสารใดบ้าง ตามสมการที่ 2.5 จะเห็นว่าค่าความสำคัญของ Term จะ

ถูกลดทอนลงและมีค่าเข้าใกล้ศูนย์ เมื่อคำนั้นปรากฏอยู่ในทุกเอกสาร ซึ่งหมายความว่าคำนั้นจะไม่มีประโยชน์ต่อการสืบค้น

$$idf_i = log\frac{N}{n_k} \tag{2.5}$$

2.1.3 Normalized Discounted Cumulative Gain (NDCG)

เป็นการวัดประสิทธิภาพของผลลัพธ์การค้นคืนเอกสารของระบบแนะนำ ระบบค้นคืน เว็บสืบค้น และแอปพลิเคชันที่เกี่ยวข้อง โดยใช้เกรดเป็นเกณฑ์ให้คะแนนกับเอกสารที่เกี่ยวข้องและให้ความสำคัญกับเอกสารที่อยู่ในลำดับต้นๆ ตามสมการที่ 2.6

$$DCG_P = \sum_{i=1}^{P} \frac{(2^{rel_i} - 1)}{log_2(1+i)} \tag{2.6}$$

เมื่อ

P แทนจำนวนผลลัพธ์การค้นคืน

rel_i แทนคะแนนที่ได้จาก judgment ความเกี่ยวข้องระหว่างเอกสารกับคำค้น ในงานวิจัยนี้แบ่งระดับของคะแนนหรือเกณฑ์ออกเป็น 5 ระดับ (5 Point Scale) คือ 0 – 4 โดย 0 คือ เอกสารไม่มีความเกี่ยวข้องกับคำค้น และ 4 คือเอกสารมีความเกี่ยวข้องกับคำค้นมากที่สุด ตามลำดับ $log_2 i$ แทนปัจจัยที่ทำให้คะแนนของเอกสารในตำแหน่งต่างๆ ถูกลดทอนลงตามอัตราส่วน การเปรียบเทียบค่า NDCG Perfect แทนด้วย IDCG (Ideal DCG) คือค่าที่มากที่สุดที่สามารถเป็นไปได้เป็นลำดับการค้นคืนที่ผู้ใช้แต่ต้องการ และเป็นการเรียงลำดับเอกสารที่มีความเกี่ยวข้องกับคำค้นมากที่สุด ถึงน้อยที่สุด คำนวณได้ตามสมการที่ 2.7

$$NDCG_P = \frac{DCG_P}{IDCG_P} \tag{2.7}$$

2.2 งานวิจัยที่เกี่ยวข้อง

Jomsri (2011) ได้ศึกษาเรื่องวิธีการเรียงลำดับผลการค้นคืนบทความวิจัย โดยใช้ Similarity Ranking ร่วมกับเวลาการโพสต์บทความ (Posted Time) โดยเก็บรวมรวมข้อมูลจาก CiteULike ใน

แต่ละบทความวิจัยประกอบด้วยข้อมูลที่บ่งชี้ถึงความสนใจในบทความของแต่ละคน ประกอบด้วย รายชื่อนักวิจัย คำสำคัญ เวลาที่ถูกโพสต์ ปีที่ตีพิมพ์ ลำดับความสำคัญ กลุ่มของเอกสาร และข้อมูลอื่นๆ ขั้นตอนการดำเนินงาน ประกอบด้วย ข้อแรกคือเวลา (Paper Posted Time) นำบทความมาเรียงตามเวลาที่ถูกโพสต์ แล้วคำนวณหาค่าคะแนนจาก $Tr = Tr - 1 - 0.05$ เมื่อกำหนดให้ r = 0, 1, 2, …, 19 และ T0 = 1 และการเรียงลำดับจาก Similarity Ranking ร่วมกับเวลา (CSTRank) เพื่อหาค่าคะแนนและลำดับผลการค้นคืนที่ดีที่สุด เพื่อพิสูจน์วิธีการเรียงลำดับผลการค้นคืนจากวิธีดังกล่าวทดสอบโดยให้นักวิชาการและผู้เชี่ยวชาญสืบค้นข้อมูล จากผลลัพธ์การค้นคืนทั้งหมด ระบบจะแสดงเฉพาะบทความที่ได้จากแต่ละดัชนีจำนวน 20 ลำดับแรก จะแสดงชื่อ บทคัดย่อ และเนื้อหา ซึ่งแต่ละคนจะต้องให้คะแนนความเกี่ยวข้องระหว่างคำค้นกับบทความนั้น กำหนดให้ช่วงคะแนนเท่ากับ 4 ถึง 0 โดย 4 หมายถึงบทความมีความเกี่ยวข้องมากที่สุดไปถึงค่าคะแนน 0 หมายถึงบทความไม่มีความเกี่ยวข้องกับคำค้นนั้นเลย ในการทดลองกำหนดค่าน้ำหนักระหว่าง Similarity กับ Static Rank เป็น 50:50 80:20 และ 90:10 ผลลัพธ์ที่ได้จากการประเมินโดยใช้ NDCG ของเอกสาร 15 ลำดับแรก พบว่า CSTRank (90:10) มีค่า NDCG สูงสุด ทำให้สามารถสรุปผลได้ว่าเมื่อนำเวลามาเป็นพิจารณาเพิ่มเติมร่วมกับ Similarity Ranking สามารถพิสูจน์ได้ว่าการเรียงลำดับให้ผลดีขึ้นโดยนักวิจัยส่วนใหญ่ให้ความสนใจกับบทความที่มีความใหม่และเพิ่งถูกตีพิมพ์มากกว่าบทความวิจัยเก่า

Zhuang, and Cucerzan (2006) ได้ศึกษาเรื่องการเรียงลำดับผลลัพธ์การค้นคืนโดยใช้ Search Query Logs เรียกว่า Q-Rank เพื่อปรับปรุงประสิทธิภาพของการเรียงลำดับผลลัพธ์การค้นคืน มีการเพิ่มตัวแปรเพิ่มเติม ได้แก่ 1. Query Extension คือชุดของคำค้นที่เกิดจากการสืบค้นภายใต้ Session เดียวกัน แล้วมีการแยกคำต่างๆ ออกด้วยการเว้นวรรค 2. Session-adjacent Query คือคำค้นข้างเคียงเกิดจากการค้นหาหลายๆ ครั้งภายใต้ Session เดียวกัน แบ่งคำค้นออกเป็น 2 ชุด คือ Q_{next} และ Q_{prev} การให้คะแนนลำดับที่วัดจากจำนวนคำค้นที่ปรากฏอยู่ใน Query Extension และ Adjacent Query โดยข้อมูลทดสอบมาจาก MSN Search Log โดยสุ่มเลือกคำค้นจำนวน 1,000 คำค้น จำนวน 2 ชุด สำหรับการพัฒนา และสำหรับการประเมินจำนวน 2,000 คำค้น ในการทดสอบ ผู้ใช้จะต้องให้คะแนนความเกี่ยวข้อง (Relevance Rating) 6 ระดับ คือ 0 หมายถึงผลลัพธ์การค้นคืนไม่มีความเกี่ยวข้องกับคำค้นจนถึงคะแนน 5 หมายถึงผลลัพธ์การค้นคืนเกี่ยวข้องกับคำค้นมากที่สุด ตามลำดับ การวัดประสิทธิภาพจาก DCG ที่ผลลัพธ์การค้นคืนในตำแหน่งที่ n เท่ากับ 10 15 และ 20 จากจำนวนเอกสารแต่ละรอบการสืบค้น c เท่ากับ 20 30 และ 40 พบว่าการเรียงลำดับที่ให้ค่า DCG มากที่สุดที่ตำแหน่ง n เท่ากับ 10 และจำนวนเอกสารที่ c เท่ากับ 30 และมีการปรับค่าน้ำหนักค่าเฉลี่ย Q-Rank มีค่าเท่ากับ 75.8% และจะมีค่าสูงสุดที่ 78.5% เมื่อค่าน้ำหนักเท่ากับ 0 หมายถึงผลลัพธ์เกิดจาก

การค้นด้วย Adjacent Query เพียงอย่างเดียว และค่าน้ำหนักเท่ากับ 0.5 ให้ค่า DCG เพิ่มขึ้นจากค่าเฉลี่ย 6.81% เท่ากับ 76.3% เมื่อทดลองปรับค่าตัวแปร n u c และค่าน้ำหนัก เท่ากับ 10 2 30 และ 0.5 ตามลำดับ ประสิทธิภาพการเรียงลำดับที่ 81.8% DCG เพิ่มขึ้น 8.99% เมื่อนำ Q-Rank จากตัวอย่างข้อมูลที่ใช้งานจริงมาพัฒนาและประเมินผล เพื่อให้ได้ผลลัพธ์ที่ใกล้เคียงกับการใช้งานจริงมากที่สุด และยังเหมาะสมกับการนำไปประยุกต์ใช้งานระบบ Web Search ที่มีอยู่แล้วเพื่อเพิ่มประสิทธิภาพและประสิทธิผลโดยไม่กระทบกับระบบการทำงานเดิมที่ใช้อยู่ได้อีกด้วย

Choochaiwattana (2010) ได้ศึกษาเรื่องกระบวนการแนะนำบทความวิจัยโดยใช้คำสำคัญ หรือ Tag จากผู้ใช้งาน ผ่านการแชร์บทความวิจัยที่แต่ละคนสนใจ โดยข้อมูลตัวอย่างถูกเก็บรวบรวมจากเว็บ CiteULike ประกอบด้วย 110 คอมมูนิตี้ 64,449 บทความวิจัย และ 262,943 คำสำคัญ ที่เกี่ยวกับงานทางด้านวิทยาการคอมพิวเตอร์ และมีการจำแนกข้อมูลของ Tag ของผู้ใช้แต่ละคน เกิดเป็นความสัมพันธ์ของข้อมูลดังนี้ ข้อแรกคือตัวแทนของผู้ใช้ (User Profile) วิเคราะห์จาก Tag Cloud เกิดจากการแชร์และBookmark บทความวิจัยของตนเอง สามารถบ่งบอกถึงความสนใจในงานแต่ละด้าน และข้อสองคือตัวแทนของบทความวิจัย เกิดจาก Tag หรือ Keyword และคำในเนื้อหามาสร้างเป็นดัชนี เพื่อได้ตัวแทนของข้อมูลทั้งสองส่วน สามารถวัดค่าคะแนนความเหมือนหรือความสนใจของผู้ใช้แต่ละคน โดยการหาค่าระยะทางระหว่าง Cosine Similarity Score ระหว่าง User Profile กับบทความความวิจัย ถ้าค่าระยะทางที่ได้มีค่าใกล้เคียงหนึ่ง ก็สามารถสรุปได้ว่าน่าจะมีความเป็นไปได้สูงที่ผู้ใช้คนนั้นจะชื่นชอบและสนใจในบทความวิจัยที่ระบบนำเสนอ จากการทดสอบระบบ ให้สมาชิกของ CiteULike และมีการแชร์บทความวิจัยจำนวน 15 คน ให้คะแนนความพึงพอใจบทความวิจัยที่ระบบแนะนำจำนวน 10 เอกสารที่มีค่า Threshold มากกว่า 0.12 และจะต้องเป็นบทความที่ผู้ใช้คนนั้นไม่เคยแชร์หรือ Bookmark มาก่อนด้วย การแสดงจะถูกสุ่มลำดับ พบว่าค่าความถูกต้องแบ่งเป็น Recall อยู่ระหว่าง 0.43 - 1.0 และ Precision อยู่ระหว่าง 0.57 - 1.0 คิดเป็นค่าความถูกต้องเฉลี่ย 79% และ f-measure ที่ 82% ผลการทดลองที่เกิดขึ้นอยู่ภายใต้สมมติฐานว่า บทความวิจัยที่ผู้ใช้แต่ละคนแชร์คือบทความที่แต่ละคนสนใจ และจะต้องไม่มีความหลากหลายทางด้านงานวิจัยหลายๆ ด้านรวมอยู่ด้วยกัน แต่สามารถนำไปประยุกต์ใช้กับทรัพยากรเว็บประเภทอื่นๆ ได้ เช่น ระบบแชร์วิดีโอ ระบบแชร์รูปภาพ

Lee, and Brusilovsky (2010) ได้ศึกษาเรื่องความสัมพันธ์ระหว่างความสนใจของผู้ใช้ระบบกับการเชื่อมโยงระหว่างผู้ใช้ (Self-defined Connections) โดยใช้ CiteULike เป็นกรณีศึกษาเพื่อนำไปพัฒนาประสิทธิภาพของ Collaborative Filtering Recommender Systems การวิเคราะห์ข้อมูลแบ่งตามความสัมพันธ์ได้สองแบบ คือ ความสัมพันธ์แบบซึ่งกันและกัน (Reciprocal) และความสัมพันธ์แบบทางเดียว (Unidirectional) สามารถแยกออกเป็นแบบทางตรงและทางอ้อม ได้

สามแบบ ได้แก่ 1. Inlink Power เท่ากับ (A intersect B) / A 2. Outlink Power เท่ากับ (A intersect B) / B 3. Overall/Jaccard Power เท่ากับ (A intersect B)/(A Union B) ส่วนการวิเคราะห์ข้อมูลส่วนที่สองคือการหา Similarity ระหว่างผู้ใช้ ได้แก่ 1. Item-based ประเมินจากปริมาณการแชร์บทความหรือการอ้างอิงไปยังบทความอื่นๆ 2. Metadata-based ประเมินจากปริมาณการแชร์นักวิจัย (Author) 3. Tag-based ประเมินจากปริมาณการแชร์คำสำคัญ (Tag) แบ่งเป็น 2 ระดับ ได้แก่ ระดับ Micro จะคิดเฉพาะคำสำคัญที่มีการแชร์โดยผู้ใช้ที่มีความสัมพันธ์ซึ่งกันและกัน และระดับ Macro คิดจากคำสำคัญทั้งหมดที่ผู้ใช้ที่มีความสัมพันธ์กันมีการแชร์คำนั้น ในการศึกษานี้มีการคิดค่า Similarity Score โดยใช้ข้อมูลการแชร์ทั้งสามส่วนแบ่งกลุ่มตามลักษณะความสัมพันธ์ของผู้ใช้ คือ 1. Items กับ Metadata 2. Tags 3. Interest Similarity กับ Item Sharing 4. Watching กับ Watched Users และ 5. เปรียบเทียบระหว่างเครือข่ายสังคมกับ Collaborative Filtering (CF-based) พบว่าผู้ใช้งานส่วนใหญ่จะเป็นแบบไม่มีความสัมพันธ์เชื่อมโยงไปยังผู้ใช้คนอื่นทั้งแบบถูกอ้างหรืออ้างอิงไปยังผู้อื่น คือ 75% และส่วนที่เหลือจะเป็นผู้ใช้แบบที่มีความสัมพันธ์เชื่อมโยงไปยังผู้ใช้คนอื่นๆ ในผู้ใช้กลุ่มที่สองจะใช้ข้อมูลที่มีลักษณะเป็นกลางคือไม่เฉพาะเจาะจง ส่งผลให้ค่า Similarity Score ระหว่างผู้ใช้ที่อยู่ในกลุ่มที่สองมีมากกว่ากลุ่มแรก และจะลดลงเมื่อระยะห่างระหว่างความสัมพันธ์เพิ่มขึ้น และค่าที่ Similarity Score ที่ได้จาก Metadata กับ tags จะมากกว่าค่าที่ได้จาก items

Bogers, and Bosch (2008) ได้ศึกษาเรื่องระบบแนะนำบทความวิจัยทางวิทยาศาสตร์ กรณีศึกษาจาก CiteULike โดยใช้ Reference Library และทดสอบวิธีการทำ Collaborative Filtering 3 วิธี พบว่าการใช้ User-based ให้ผลลัพธ์การค้นคืนดีที่สุด ข้อมูลที่ใช้แบ่งเป็นห้าประเภท ได้แก่ 1. Topic-related Metadata คือรายละเอียดเกี่ยวกับหัวข้องานวิจัย เช่น ชื่อบทความ และรายละเอียดการตีพิมพ์ 2. Personal-related Metadata รายละเอียดเกี่ยวกับบุคคล เช่น รายชื่อนักวิจัย (Authors) พิมพ์ 3. Temporal Metadata เช่น ปีที่ตีพิมพ์ เดือน 4. Miscellaneous Metadata เช่น ประเภทของบทความวิจัย หรือรายละเอียดอื่นๆ เช่น รายละเอียดสำนักพิมพ์ Volume Number จำนวนหน้า DOI และ ISSN/ISBN และ URL ที่ลิงค์ไปยังบทความวิจัยฉบับเต็ม 5. User-Specific Metadata เป็นรายละเอียดที่ผู้ใช้เป็นผู้กำหนด เช่น คำสำคัญ ความคิดเห็น (Comment) และลำดับความสำคัญ (Priority Reading) ศึกษาวิธีการสร้างรายการบทความวิจัยที่เกี่ยวข้องกันจาก User's Reference Library เรียกว่า "การหาสิ่งที่ดีที่สุด" (Find Good Item) ผู้ใช้จะถูกนำเสนอบทความโดยใช้พื้นฐานจากข้อมูลส่วนตัว

Seki, Qin, and Uehara (2010) ได้ศึกษาเรื่องผลกระทบของ Social Bookmarks หรือ Social Tags ในการค้นคืนข้อมูลบรรณานุกรม Genomics Track เฉพาะบทความที่มีการอ้างอิงถึง

CiteULike คือจะต้องมี Social Tag อย่างน้อยหนึ่งคำ คิดเป็นหนึ่งในสี่ของบทความทั้งหมด ข้อมูลชื่อบทความวิจัย บทคัดย่อ และ หัวข้อทางการแพทย์ (Medical Subject Heading: MeSH) จากฐานข้อมูล TREC ถูกนำมาสร้างเป็นดัชนีสำหรับทดสอบ ดัชนีแบ่งเป็น 4 ชุดการทดสอบ ได้แก่ 1. None สร้างจากชื่อบทความและบทคัดย่อ 2. MeSH สร้างจาก None Index และ Medical Subject Heading หรือ MeSH terms 3. CiteULike สร้างจาก None Index และ CiteULike Tags 4. Both สร้างจาก 1. 2. และ 3. รวมกัน สำหรับการวัดคุณภาพของบทความวิจัยจะใช้วิธีการนับจำนวนคำสำคัญที่ถูกใช้ซ้ำกันในแต่ละเอกสาร พบว่าค่า MAP มีค่าสูงขึ้นเมื่อ Threshold เท่ากับ 2 และลดลงเมื่อ Threshold มีค่าเพิ่มขึ้น การวัดคุณภาพของคำสำคัญจาก Inverse Document Frequency (*idf*) เท่ากับ *log(N/DF)* เมื่อ *N* แทนจำนวนเอกสารทั้งหมด และ *df* แทนจำนวนเอกสารที่คำสำคัญนั้นปรากฏอยู่ พบว่าค่า MAP จะมีค่าสูงสุด ควรให้ Threshold มีค่าอยู่ระหว่าง 7.7 ถึง 9.2 การวัดคุณภาพของคำสำคัญ แยกตามปีของบทความวิจัย เปรียบเทียบระหว่างดัชนีที่มีคำสำคัญกับไม่มี พบว่าเมื่อปริมาณคำที่เพิ่มขึ้นในแต่ละปีส่งผลให้ค่า MAP เพิ่มขึ้นอย่างต่อเนื่องตามลำดับ

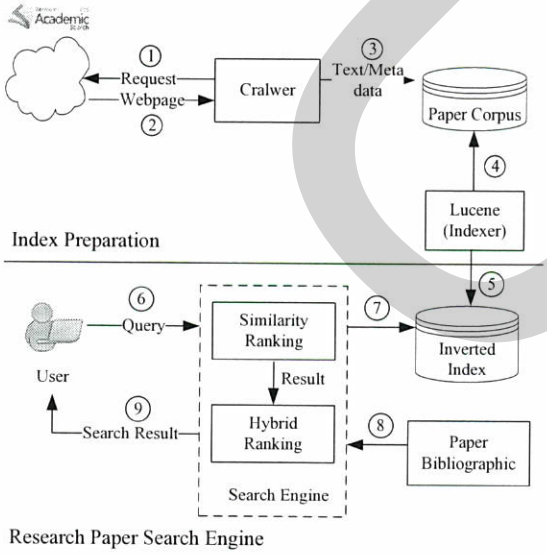
จากงานวิจัยที่เกี่ยวข้องมีการนำข้อมูลทางบรรณานุกรมต่างๆ เช่น เวลา คำสำคัญ คำค้นข้างเคียง มาช่วยเสริมให้ผลลัพธ์การค้นคืนมีประสิทธิภาพและประสิทธิผลดีขึ้น ทั้งในส่วนของระบบค้นคืนและระบบแนะนำ ในการวิจัยนี้จึงนำเอาแนวความคิดที่ได้มาประยุกต์ใช้เพิ่มเติมดังกล่าวมาสร้างเป็นตัวแบบสำหรับระบบค้นคืนบทความวิจัยในบทถัดไป

บทที่ 3
ระเบียบวิจัย

ในบทนี้จะกล่าวถึงขั้นตอนการดำเนินงานวิจัย โดยจะแบ่งออกเป็นสองส่วนคือ ขั้นตอนการดำเนินการวิจัย จะเริ่มตั้งแต่การเก็บรวบรวมข้อมูล การวิเคราะห์จัดเก็บข้อมูลกลุ่มประชากร การสร้างดัชนี การสร้างต้นแบบ การออกแบบทดสอบ และการประเมินผล รวมถึงเครื่องมือที่ใช้ โดยมีรายละเอียดดังนี้

3.1 ทฤษฎีการวิเคราะห์ปัญหาและศึกษาค้นคว้าข้อมูล

วิธีการดำเนินการวิจัยเป็นวิธีการสร้างตัวแบบสำหรับระบบค้นคืนมีขั้นตอนต่างๆ ตั้งแต่การเก็บรวบรวมข้อมูล การวิเคราะห์ข้อมูล การสร้างดัชนี การสร้างตัวแบบ และระบบค้นคืนสำหรับทดสอบตัวแบบ แล้วนำผลได้ที่ไปประเมินผลในขั้นตอนสุดท้าย แสดงดังภาพที่ 3.1 มีรายละเอียดดังนี้



ภาพที่ 3.1 ขั้นตอนการทำงานของระบบค้นคืนเอกสาร

3.1.1 การเก็บรวบรวมและการวิเคราะห์ข้อมูล (Data Preparation)

กระบวนการเตรียมข้อมูล สำหรับนำเข้าในขั้นตอนถัดไปคือการสร้างดัชนีของระบบค้นคืน มีขั้นตอนดังนี้

3.1.1.1 การครอลว์ข้อมูล (Crawl)

Crawler ทำหน้าที่อ่านข้อมูลจาก http://academic.research.microsoft.com เป็นเว็บที่ให้บริการข้อมูลผลงานวิจัยที่ตีพิมพ์ในวารสารและงานประชุมวิชาการ ลักษณะของข้อมูลในเว็บไซต์ที่ได้ จะมีลักษณะเป็นสตรีมไบต์ นำเข้ากระบวนการวิเคราะห์คำต่อไป เพื่อตัดเอาเฉพาะเนื้อหาสำคัญ ตัวอย่างเว็บและข้อมูลตามภาคผนวก ก

3.1.1.2 การวิเคราะห์และการคัดกรองข้อมูล (Parsing)

ข้อมูลที่ได้จากการอ่านจาก Crawler มีรูปแบบโครงสร้างอยู่ในลักษณะของ HTML Element หรือ HTML Tags โดย Parser วิเคราะห์ว่าข้อมูลที่ต้องการอยู่ภายใต้ Element ใด เพื่อสกัดข้อมูลเฉพาะสาระสำคัญที่ต้องการออกมาเท่านั้น การสกัดแต่ละครั้งจะถูกเก็บลงฐานข้อมูล รายละเอียดการออกแบบฐานข้อมูลเพิ่มเติมในภาคผนวก ข

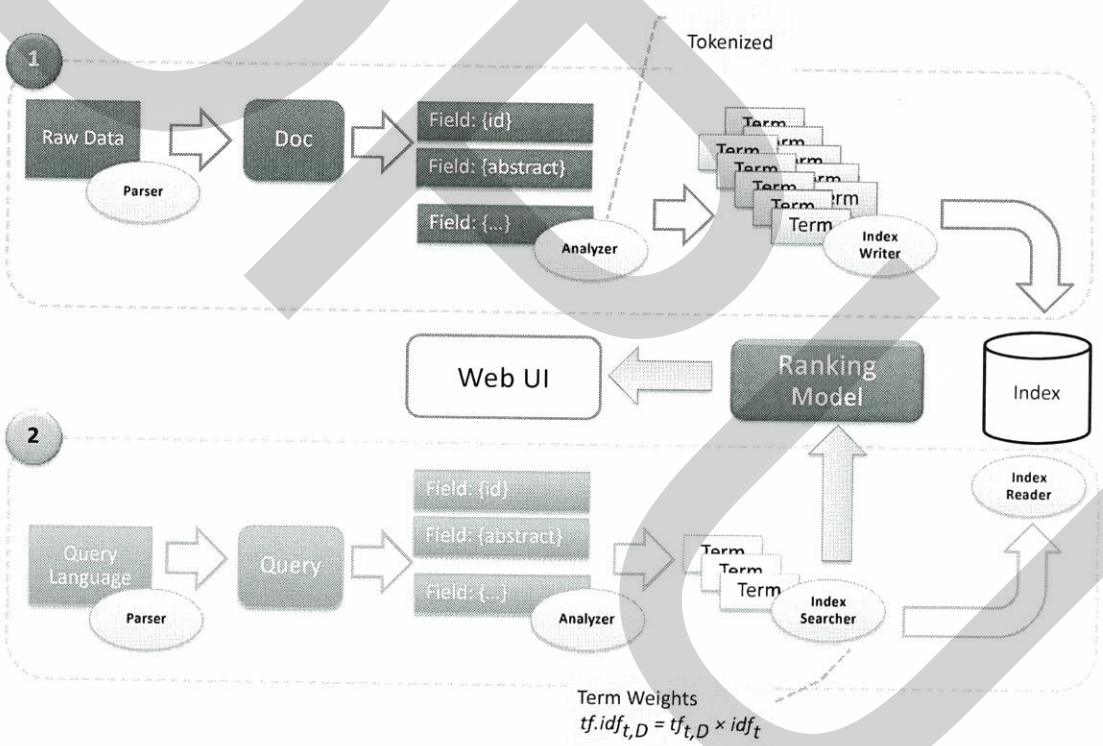
จากขั้นตอนข้างต้น การรวบรวมข้อมูลดำเนินการระหว่าง เดือนมิถุนายน - สิงหาคม พ.ศ. 2556 ประกอบ ด้วยบทความวิจัยจำนวน 71,828 บทความ สามารถจำแนกเป็นแต่ละประเภทได้ตามตารางที่ 3.1 ในแต่ละบทความประกอบด้วยข้อมูลทางบรรณานุกรม ได้แก่ ชื่อบทความวิจัย รายชื่อนักวิจัย บทคัดย่อ คำสำคัญ จำนวนบทความวิจัยอ้างอิง (Reference) จำนวนผลงานถูกอ้างอิง (Citation) ปีและรายชื่อแหล่งตีพิมพ์ ประเภทการตีพิมพ์บทความวิจัย ปีที่จัดพิมพ์ จำนวนบทความวิจัยที่เคยตีพิมพ์มาแล้วทั้งหมด จำนวนบทความที่อ้างอิงถึง และข้อมูลอื่นๆ ซึ่งข้อมูลแต่ละส่วนจะถูกแยกเก็บ เรียกว่า ฟิลด์ (Field) รายละเอียดเพิ่มเติมและตัวอย่างข้อมูล ตามภาคผนวก ง

ตารางที่ 3.1 คลังเอกสาร

คลังเอกสาร	วารสารวิชาการ (Journal)	การประชุมวิชาการ (Conference)	รวม
บทความวิจัย	28,320	43,508	71,828
แหล่งตีพิมพ์	4,283	2,885	7,168

3.1.2 การสร้างดัชนี

ดัชนีคือโครงสร้างข้อมูลที่แปลงจากเอกสารบทความวิจัย เพื่อให้ระบบค้นคืนสามารถเข้าถึงข้อมูลและค้นหาได้อย่างรวดเร็ว และนำมาใช้เป็นฐานข้อมูลเอกสารหรือคลังเอกสาร ซึ่งในขั้นตอนของการสร้างดัชนีในการวิจัยนี้ มีการนำไลบรารีลูซีน (Lucene) หรือเรียกว่า Standard Analyzer เป็นเครื่องมือที่ช่วยทำหน้าวิเคราะห์เอกสาร ข้อความ และคำที่อยู่ในบทความวิจัยจำแนกข้อมูลที่ได้จากกระบวนการวิเคราะห์คำออกเป็นฟิลด์ จากภาพที่ 3.2 ข้อมูลแต่ละฟิลด์จะถูกนำเข้ามาเพื่อผ่านกระบวนการตัดคำ (Tokenized) จัดเก็บลงในคลังเอกสารรูปแบบของ Inverted Index ทำให้การค้นคืนมีประสิทธิภาพและยืดหยุ่นมากยิ่งขึ้น เพื่อเข้าสู่กระบวนการสร้างส่วนติดต่อกับผู้ใช้และสร้างตัวแบบสำหรับการทดสอบ และประเมินผลในขั้นตอนถัดไป



ภาพที่ 3.2 ขั้นตอนการวิเคราะห์คำเพื่อสร้างและค้นคืนผ่านดัชนี

ในส่วนแรกเป็นการสร้างดัชนีข้อมูลนำเข้าจะได้จากฐานข้อมูลทีละเอกสาร และแยกออกเป็นฟิลด์ เพื่อแบ่งแยกข้อมูลออกเป็นหมวดหมู่ที่ชัดเจน จากนั้น Analyzer จะนำเอกสารมาตัดคำ (Tokenized) เพื่อคำนวณหาค่า Term Weight ให้ Index Writer เขียนลงในคลังเอกสาร ข้อมูลแต่ละฟิลด์แสดงตามตารางที่ 3.2 ส่วนที่สองเป็นการอ่านหรือการค้นคืน คำที่ได้จากผู้ใช้งานจะต้อง

ผ่านกระบวนการเช่นเดียวกับการนำเข้า แต่สิ่งที่ได้จาก Index Reader คือรายการของเอกสารที่ Hit กับคำค้น เข้าสู่ตัวแบบเพื่อเรียงลำดับผลลัพธ์การค้นคืนใหม่ อธิบายการคำนวณในหัวข้อถัดไป

ตารางที่ 3.2 ฟิลด์ข้อมูลที่ใช้ทำดัชนี

ลำดับ	ฟิลด์	รายละเอียด	ดัชนี	ประเภท
1	ArticleId	หมายเลขบทความวิจัย	Not Analyzed	Numeric
2	Title	ชื่อบทความวิจัย	Tokenized	String
3	Tags	รายการคำสำคัญ	Tokenized	String
4	Abstracts	บทคัดย่อ	Tokenized	String
5	Author	รายชื่อผู้แต่ง	Tokenized	String
6	Year	ปีที่ปีพิมพ์	Not Analyzed	Numeric
7	CitationContext	จำนวนการถูกอ้างอิง	Not Analyzed	Numeric
8	Reference	จำนวนการอ้างอิง	Not Analyzed	Numeric
9	CitationList	จำนวนการถูกอ้างอิงและปรากฏในเนื้อหา	Not Analyzed	Numeric
10	ArticleUrl	URL ของบทความวิจัย	Not Analyzed No Norms	String
11	PublishId	หมายเลขแหล่งตีพิมพ์	Not Analyzed	Numeric
12	PublishTitle	ชื่อแหล่งตีพิมพ์	Tokenized	String
13	PublishType	ประเภทของแหล่งตีพิมพ์	Tokenized	String
14	FieldOfStudy	กลุ่ม	Tokenized	String
15	Publications	จำนวนบทความวิจัยที่ตีพิมพ์มาแล้วทั้งหมด	Not Analyzed	Numeric
16	CitationCount	จำนวนการถูกอ้างอิง	Not Analyzed	Numeric
17	SelfCitation	จำนวนการอ้างอิงภายในแหล่งตีพิมพ์เดียวกัน	Not Analyzed	Numeric
18	PublisherUrl	URL ของแหล่งตีพิมพ์	Not Analyzed No Norms	String
19	YearRange	อายุของแหล่งตีพิมพ์	Not Analyzed No Norms	Numeric

3.1.3 การสร้างตัวแบบ (Hybrid Model)

จากขั้นตอนที่ 3.1.2 ระบบจะได้รายการของบทความวิจัยของแต่ละดัชนีออกมา ดัชนีที่ได้จาก Similarity Model มี 2 แบบคือ Full Text Index หรือ Similarity Ranking เรียกว่า Similarity1 และ Similarity Ranking เพิ่มค่าน้ำหนักให้กับฟิลด์ในเอกสาร (Field Boost) คือ ชื่อบทความวิจัย (Title) บทคัดย่อ (Abstract) และ คำสำคัญ (Keyword) ด้วยค่าน้ำหนักแต่ละตัวเป็น 3 2 และ 1 ตามลำดับ เรียกว่า Similarity2 สามารถคำนวณได้จากสมการที่ 3.1 ซึ่งค่าคะแนนที่ได้หมายถึง Similarity Measure ระหว่างแต่ละ Term ใน Query เทียบกับแต่ละเอกสารและ Hybrid Model มี 2 แบบ โดยนำ Similarity Feature ผสมกับข้อมูลทางบรรณานุกรม (Bibliographic) ที่ได้จาก Similarity Model คือ Similarity1 กับผสมกับข้อมูลบรรณานุกรม เรียกว่า Hybrid1 และ Similarity2 ผสมกับข้อมูลบรรณานุกรม เรียกว่า Hybrid2

$$Sim(q,d) = \sum_{t\ in\ q} (tf(t\ in\ d) \times idf(t)^2 \times b(t.field\ in\ d) \times ln(q)) \times c(q,d) \times qN(q) \quad (3.1)$$

เมื่อ

$tf(t\ in\ d)$ แทน Term Frequency

$idf(t)$ แทน Inverse Document Frequency

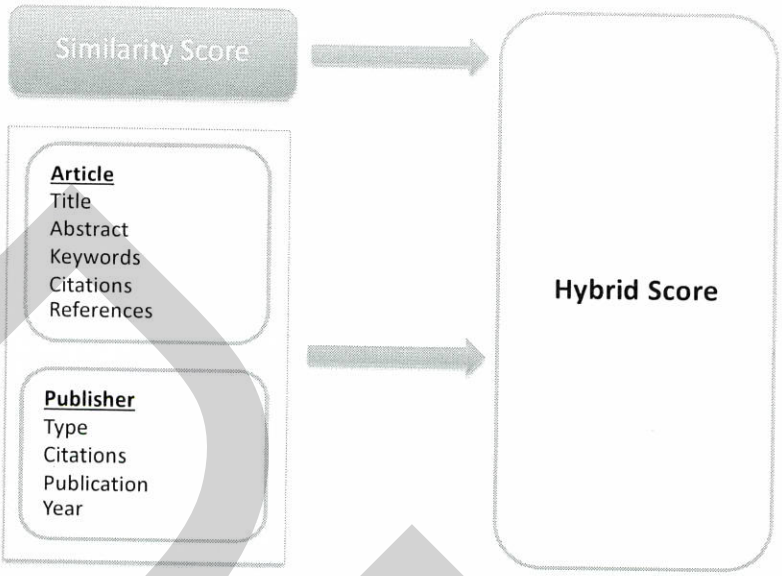
$b(t.field\ in\ d)$ แทน Field Boost และ Document Boost

$ln(q)$ แทน $lenNorm$ คือ Normalized ของ Field คิดจากจำนวน Term ใน Field

$c(q,d)$ แทน $coord(q,d)$ จำนวน Term ใน Query ที่ปรากฏในเอกสาร

qN แทน $queryNorm(q)$ ค่า Normalized ของคะแนนแต่ละ Query Term

ข้อมูลบรรณานุกรมแบ่งออกเป็น 2 ส่วนคือข้อมูลบทความวิจัย เรียกว่า Article หรือ Paper Quality และข้อมูลแหล่งตีพิมพ์ การตีพิมพ์บทความวิจัย เรียกว่า Publisher Quality เมื่อพิจารณาถึงฟิลด์ข้อมูลที่เป็นตัวแปรสำคัญ ตามภาพที่ 3.3



ภาพที่ 3.3 กรอบแนวคิดการสร้างตัวแบบ

จากภาพที่ 3.3 กำหนดให้ความสัมพันธ์ระหว่าง Similarity Feature กับ Bibliographic Feature ของเอกสารงานวิจัยและแหล่งตีพิมพ์เพื่อคำนวณค่า Hybrid Score ตามสมการที่ 3.2

$$Hybrid\ Score\ = Sim(\propto) + Bib(1-\propto) \tag{3.2}$$

เมื่อ

$\propto$ แทน ค่าน้ำหนัก

Sim แทน Similarity Score ของคำค้น

Bib แทน Bibliographic Score คิดจากค่าเฉลี่ยจากการวัดค่าคุณภาพของบทความวิจัยร่วมกับคุณภาพของผู้จัดพิมพ์งานวิจัย ได้จากสมการที่ 3.3 และ 3.4 ตามลำดับ

$$QA\ =\ R(\beta)\ +\ CA(1-\beta) \tag{3.3}$$

เมื่อ

QA แทนคุณภาพของเอกสารงานวิจัย (Article Quality)

β เท่ากับ 0.9 ค่าน้ำหนัก

R แทนจำนวนเอกสารอ้างอิงภายในบทความวิจัย ค่าที่นำมาใช้อยู่ในรูปแบบการทำ Scale Normalized อยู่ระหว่าง 0 ถึง 1 คำนวณจากเอกสาร 30 เอกสารแรกที่ได้จากการค้นคืน

CA แทนจำนวนเอกสารที่มีการอ้างอิงถึงบทความวิจัยนั้น ค่าที่นำมาใช้ทำ Scale Normalized อยู่ระหว่าง 0 ถึง 1 คำนวณจากเอกสาร 30 เอกสารแรกที่ได้จากการค้นคืน

$$QP = Type \times CP \tag{3.4}$$

เมื่อ

QP แทนคุณภาพของผู้จัดพิมพ์งานวิจัย (Publisher Quality)

$Type$ แทนประเภทของผู้จัดพิมพ์ กำหนดให้วารสารวิชาการ (Journal) เท่ากับ 1.0 งานประชุมวิชาการ (Conference) เท่ากับ 0.1

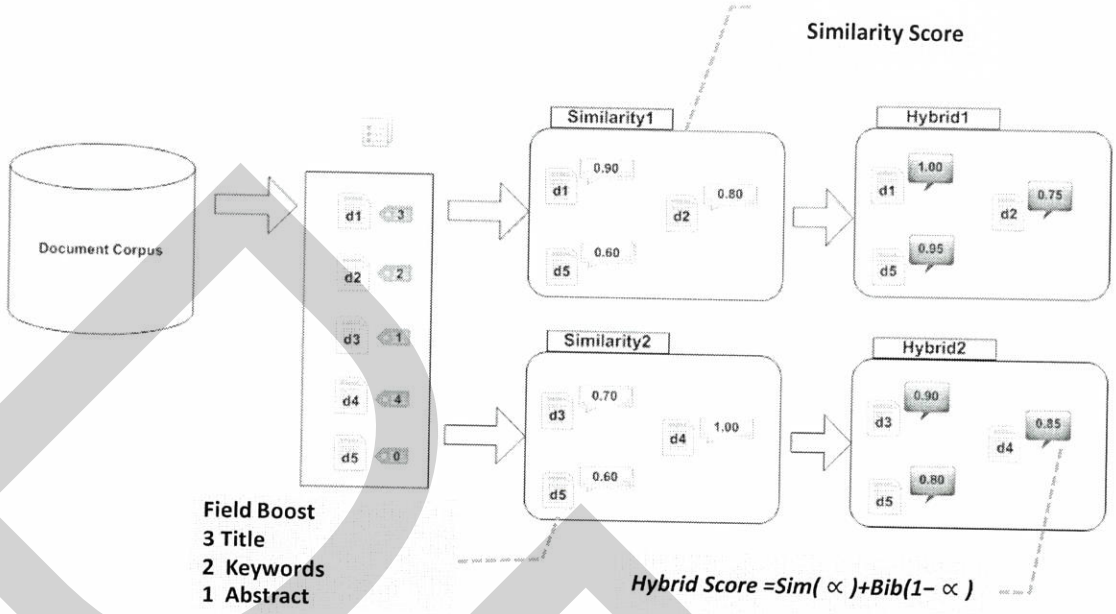
CP แทนจำนวนเอกสารที่มีการอ้างอิงถึงผู้จัดพิมพ์ เปรียบเทียบกับจำนวนเอกสารที่ตีพิมพ์ของผู้จัดพิมพ์นี้ และค่าที่ได้มาทำ Scale Normalized อยู่ระหว่าง 0 ถึง 1

ดังนั้นสามารถสรุปตัวแบบได้ตามตารางที่ 3.3 และภาพที่ 3.4

ตารางที่ 3.3 คุณลักษณะของตัวแบบ

Index	Similarity Feature	Boost Field	Bibliographic Feature
Similarity1	✓	-	-
Similarity2	✓	✓	-
Hybrid1	✓	-	✓
Hybrid2	✓	✓	✓

ตัวอย่างการคำนวณ Hybrid Score ตามภาพที่ 3.4 จากตัวแบบที่สร้างขึ้นบทความวิจัยจะมีค่าคะแนนและการเรียงลำดับผลลัพธ์การค้นคืนใหม่ด้วยเช่นกัน



ภาพที่ 3.4 การคำนวณ Hybrid Score

3.1.4 การทดลอง

การทดสอบเพื่อพิสูจน์ตัวแบบที่สร้างขึ้นจะสามารถเรียงลำดับผลลัพธ์การค้นคืนที่ดีขึ้นตามสมมติฐาน จึงจัดทำระบบ Paper Search Engine เป็นหน้าเว็บ GUI ให้ผู้ใช้ติดต่อกับระบบค้นคืนในการทดสอบตัวแบบ โดยเชิญนักศึกษาระดับปริญญาโท ปริญญาเอก อาจารย์และนักวิจัยด้านวิทยาการคอมพิวเตอร์ภายใต้ขอบเขตเนื้อหาของการวิจัยนี้ โดยกำหนดให้ผู้ทดสอบแต่ละคนใส่คำค้นที่ต้องการเป็นคำ หรือประโยคใดๆ ก็ได้ในหน้าเว็บ ระบบจะสืบค้นข้อมูลจากดัชนีจาก Similarity1 และ Similarity2 เพื่อคำนวณหาค่า Similarity Score และนำคะแนนที่ได้นำเข้าสู่ตัวแบบเพื่อเข้าสู่กระบวนการประมวลผลภายใต้ตัวแบบเพื่อหาค่า Hybrid Score อีกครั้ง โดยก่อนที่จะแสดงผลให้ผู้ทดสอบประเมิน ระบบจะตรวจสอบเอกสารที่ได้ในแต่ละดัชนีที่เป็นเอกสารเดียวกัน ระบบจะรวมผลลัพธ์ให้เหลือเอกสารเพียงเอกสารเดียวเพื่อไม่ให้มีการแสดงผลบนหน้าจอซ้ำ และระบบจะแสดงผลแบบสุ่มลำดับเพื่อมิให้ผู้ทดสอบเกิดความลำเอียงในการให้คะแนน ตัวอย่างหน้าจอในภาคผนวก ค

หน้าเว็บที่แสดงผลลัพธ์ จะแสดงรายละเอียดบทความวิจัย ได้แก่ ชื่อบทความวิจัย และบทคัดย่อ โดยผู้ทดสอบจะต้องอ่านรายละเอียดทั้งสองส่วน แล้วให้คะแนนบทความที่กำลังพิจารณา

ว่ามีความเกี่ยวข้องกับคำค้นมากน้อยแค่ไหน ซึ่งคะแนนที่ได้ดังกล่าวจะนำไปประเมินผลตัวแบบการเรียงลำดับ โดยการทดสอบมีขั้นตอนดังนี้

1. ผู้ทดสอบระบุคำค้นที่ต้องการในหน้าเว็บ

2. ระบบจะค้นคืนเอกสาร 30 ลำดับแรก ของแต่ละดัชนี โดยระบบจะตรวจสอบเอกสารที่แสดงผลซ้ำ และสุ่มลำดับการแสดงผลบนหน้าเว็บ เพื่อไม่ให้ผู้ทดสอบเกิดความลำเอียงในการให้คะแนนในแต่ละเอกสาร

3. ผู้ทดสอบให้คะแนนเอกสาร (Judgment Score) แต่ละเอกสารมีความเกี่ยวข้องกับคำค้นอย่างไร คะแนนอยู่ระหว่าง 4 ถึง 0 มีความหมาย ตามตารางที่ 3.4

4. ระบบบันทึกข้อมูล

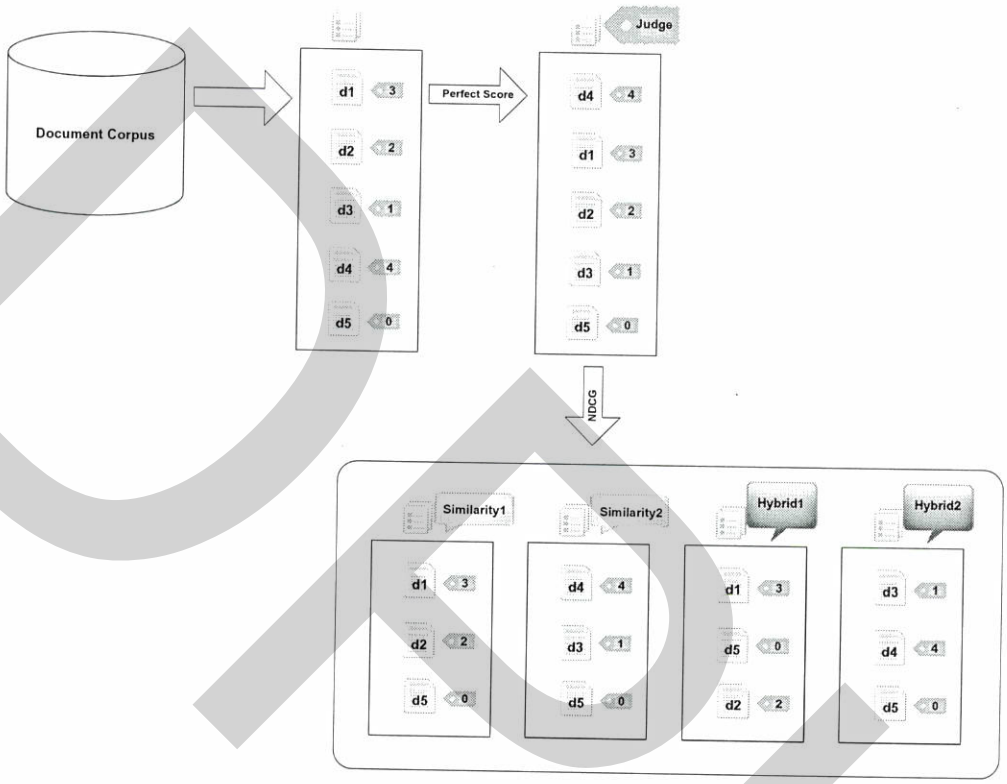
ตารางที่ 3.4 Judgments Score

คะแนน	คำอธิบาย
4	มีความเกี่ยวข้องกันอย่างมาก (Very Relevant)
3	มีความเกี่ยวข้อง (Relevant)
2	มีความเกี่ยวข้องกันบางส่วน (Somewhat Relevant)
1	มีความเกี่ยวข้องกันเป็นส่วนน้อย (Only Slightly Relevant)
0	ไม่มีความเกี่ยวข้องกัน (Non-Relevant)

3.1.5 การประเมินผล

เมื่อได้ Judgment Score ของบทความวิจัยถูกนำมาประเมินผล 2 แบบ คือแบบแรกคิดค่า NDCG กล่าวไปแล้วในบทที่ 2 เป็นการประเมินว่าลำดับผลลัพธ์การค้นคืนที่ได้มีประสิทธิผลเป็นอย่างไร โดยนำเอกสารทั้งหมดเรียงลำดับตาม Judgment Score เพื่อหา DCG Perfect หรือ Ideal DCG และกลุ่มของบทความวิจัยแยกตามดัชนีเพื่อหา NDCG ตามภาพที่ 3.6 ส่วนที่สองคือการวัดประสิทธิภาพคือการหาค่าเฉลี่ยความถูกต้อง เรียกว่า Mean Average Precision (MAP) เป็นการประเมินว่าเอกสารที่ได้จากการค้นคืนถูกต้องตรงกับความต้องการของผู้ใช้มากน้อยแค่ไหน จะตัดคะแนนความถูกต้องจาก 0 ถึง 4 ด้วยค่าคะแนนเท่ากับ 3 ถ้าเอกสารที่ได้คะแนนเท่ากับ 0 ถึง 2

หมายถึงเอกสารนั้นไม่เกี่ยวข้องกับคำค้น และคะแนนเท่ากับ 3 ถึง 4 หมายถึงเอกสารมีความเกี่ยวข้องกับคำค้น



ภาพที่ 3.5 ขั้นตอนการประเมินผล

3.2 เครื่องมือที่ใช้ในการวิจัย

วิธีการดำเนินการวิจัยเป็นวิธีการสร้างระบบตัวแบบสำหรับ Search Engine มีขั้นตอนต่างๆ ตั้งแต่การเก็บรวบรวมข้อมูล การวิเคราะห์ข้อมูล การสร้างดัชนี การสร้างตัวแบบ และระบบค้นคืนสำหรับทดลองตัวแบบ มีเครื่องมือที่ใช้อำนวยความสะดวกในการวิจัยดังนี้

3.2.1 Crawler เป็นโปรแกรมที่พัฒนาด้วย Java Application เพื่ออ่านข้อมูลที่บนเว็บ และจัดเก็บข้อมูลที่ต้องการลงในฐานข้อมูล

3.2.2 MySql เป็นฐานข้อมูลสำหรับเก็บข้อมูลที่ได้จากการวิเคราะห์คำ

3.2.3 Lucene เป็นจาวาไลบรารีสำหรับการสร้างดัชนีเพื่อใช้เป็นคลังเอกสาร

บทที่ 4
ผลการดำเนินงาน

จากการดำเนินการวิจัยในบทที่ 3 จะได้ผลการทดสอบจากผู้ร่วมทดสอบ 20 คน แต่ละคนประเมินตั้งแต่ 1 ถึง 3 ครั้ง รวมจำนวนคำค้นทั้งหมด 37 ครั้ง บทความที่มี Ranking Score อยู่ใน 30 ลำดับแรกในแต่ละดัชนีจะถูกค้นคืนออกมา แล้วนำมาตรวจสอบก่อนเพื่อไม่ให้แสดงผลซ้ำกัน และเรียงลำดับใหม่เพื่อไม่ให้เกิดการเอนเอียงระหว่างการให้คะแนน ใช้วิธีการประเมินออกเป็น 2 แบบคือ การหาค่า Normalized Discount Cumulative Gain (NDCG) และการประเมินอีกวิธีหนึ่งคือค่า Mean Average Precision (MAP) เป็นการวัดประสิทธิภาพโดยการค่าเฉลี่ยความถูกต้องของผลลัพธ์การค้นคืนเอกสาร ว่าเอกสารที่เกี่ยวข้องกับคำค้นถูกค้นคืนตามที่ต้องการหรือไม่ จากการสร้างตัวแบบทั้ง 4 วิธี ได้ผลดังนี้

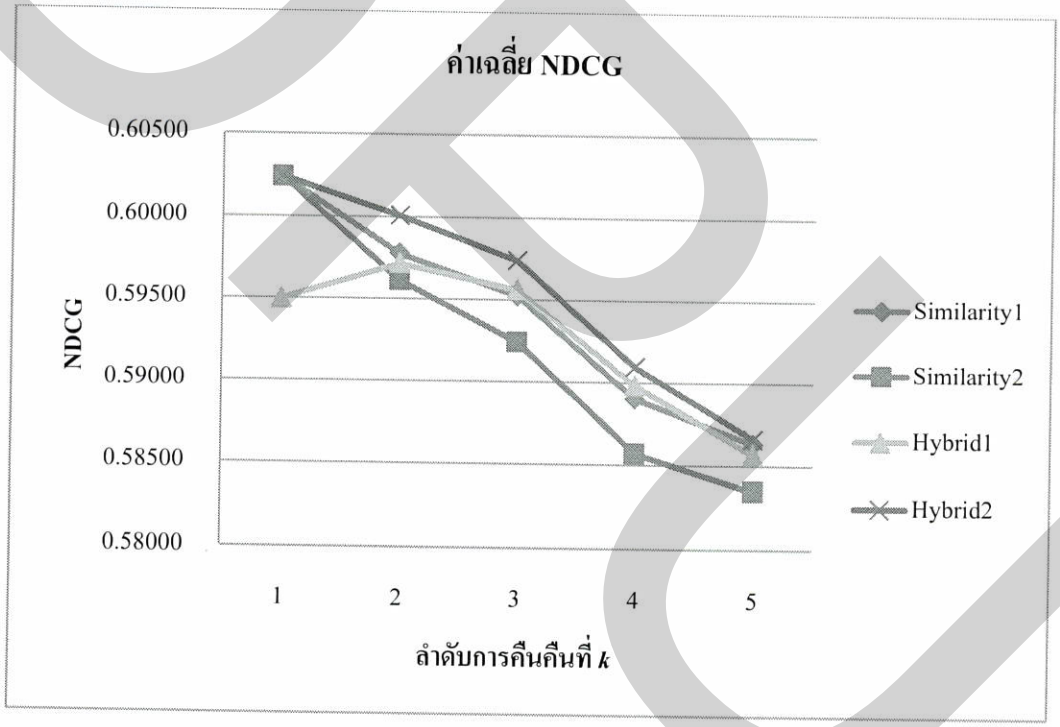
4.1 ค่าเฉลี่ย NDCG

เป็นการวัดประสิทธิผลการเรียงลำดับผลลัพธ์การค้นคืนที่เป็นมาตรฐานที่นิยมใช้ทั่วไปในระบบค้นคืน โดยในการวิจัยนี้ผู้ร่วมทดสอบจะให้คะแนนความเกี่ยวข้อง (Relevant) ระหว่างคำค้นกับบทความวิจัย

จากตารางและกราฟที่ 4.1 แสดงค่าเฉลี่ย NDCG ที่ได้จาก Judgment Score ของผลลัพธ์การค้นคืนใน 5 ลำดับแรก พบว่าที่ตำแหน่ง k เท่ากับ 1 ดัชนีที่มาจาก Similarity1 กับ Hybrid2 ให้ค่าเฉลี่ยเท่ากันคือ 0.60238 และเมื่อพิจารณาผลใน 4 ลำดับถัดไปพบว่า Hybrid2 ได้ผลการประเมินสูงสุดเมื่อเทียบกับดัชนีอื่นๆ ค่าดัชนีของ Hybrid1 มีการกำหนด $\propto$ เท่ากับ 0.9 ซึ่งเป็นอัตราส่วนระหว่างค่าน้ำหนักของ Similarity Ranking กับ Static Ranking ที่มีส่วนช่วยให้อัตราการเรียงลำดับผลลัพธ์การค้นคืนดีขึ้น เมื่อเปรียบเทียบกับการเรียงลำดับบทความวิจัย ซึ่งส่วนใหญ่จะให้ความสำคัญกับความเหมือนมากกว่าคุณภาพของเอกสาร

ตารางที่ 4.1 ค่าเฉลี่ย NDCG

k	Similarity1	Similarity2	Hybrid1	Hybrid2
1	0.60238	0.60238	0.59497	0.60238
2	0.59776	0.59613	0.59715	0.60007
3	0.59524	0.59240	0.59561	0.59739
4	0.58898	0.58567	0.58980	0.59093
5	0.58630	0.58349	0.58565	0.58660



ภาพที่ 4.1 ค่าเฉลี่ย NDCG

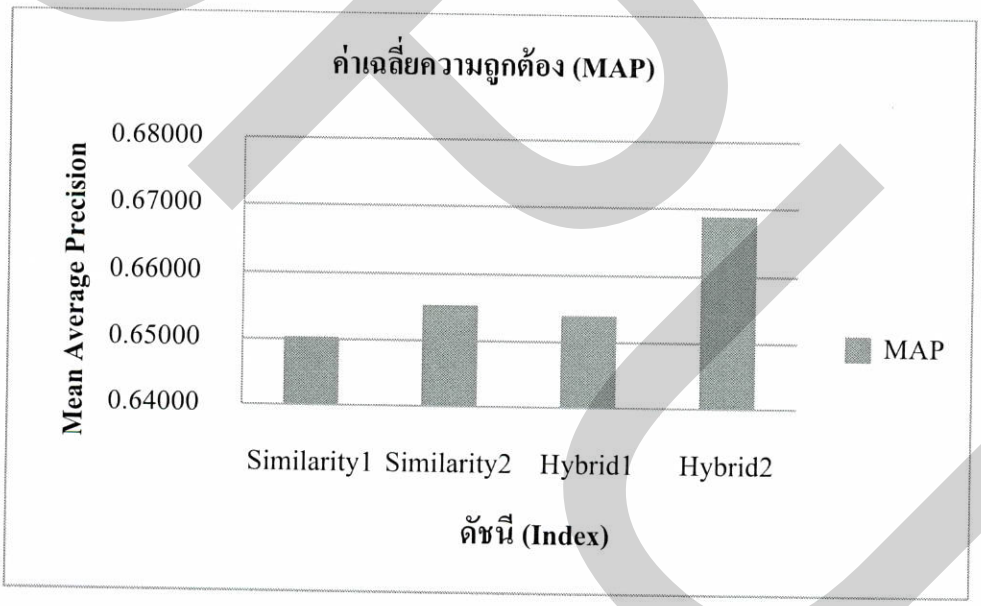
4.2 ค่าเฉลี่ย MAP

การหาค่าเฉลี่ยความถูกต้อง (Mean Average Precision: MAP) เป็นการวัดประสิทธิภาพ โดยคิดจาก Judgment Score ของบทความวิจัยที่ได้จากการทดสอบ ในการวิจัยนี้ตัดช่วงของ Judgment Score

ที่ 3 หมายถึงค่าคะแนนระหว่าง 3-4 ถือว่าบทความวิจัยนั้นเกี่ยวข้องกับคำค้น และ ระหว่าง 0-2 ถือว่าเอกสารนั้นไม่มีความเกี่ยวข้องกับคำค้น

ตารางที่ 4.2 ค่าเฉลี่ย MAP

ดัชนี	ค่าเฉลี่ย
Similarity1	0.65030
Similarity2	0.65537
Hybrid1	0.65394
Hybrid2	0.66881



ภาพที่ 4.2 ค่าเฉลี่ย MAP

จากตารางที่ 4.2 และภาพที่ 4.2 พบว่า เมื่อตัดช่วง Judgment Score ที่ 3 พบว่าผลลัพธ์จากดัชนีของ Hybrid2 สามารถให้ค่าความถูกต้องสูงสุด และ Similarity2 Hybrid1 Similarity1 ให้ค่าความถูกต้องลดลงตามลำดับ

บทที่ 5
สรุป อภิปรายผล และข้อเสนอแนะ

วัตถุประสงค์ของการวิจัยครั้งนี้ เพื่อสร้างตัวแบบสำหรับการเรียงลำดับผลลัพธ์การค้นคืนบทความวิจัยบนระบบค้นคืน โดยมีพื้นฐานบนสมติฐานและแนวคิดจากการสืบค้นหาบทความวิจัยในปัจจุบัน ที่ส่วนใหญ่มักจะทำผ่านระบบฐานข้อมูลหรือการเปรียบเทียบความคล้ายคลึงระหว่างบทความวิจัยกับคำค้น โดยใช้หลักการของ Similarity ดังนั้นเพื่อให้ระบบค้นคืนที่สามารถค้นคืนบทความวิจัยที่มีประสิทธิผลมากขึ้นควรนำปัจจัยอื่นๆ มาพิจารณาเพิ่มเติม เช่น ปริมาณบทความวิจัยที่อ้างอิงถึง ปริมาณการเอกสารอ้างอิงของบทความวิจัย เรียกเนื้อหาส่วนนี้ว่าเป็นคุณภาพของบทความวิจัย และอีกส่วนหนึ่งคือคุณภาพของแหล่งตีพิมพ์ เนื้อหาที่นำมาพิจารณา ได้แก่ ปริมาณการอ้างอิงถึง ช่วงเวลาการตีพิมพ์ ประเภทของแหล่งตีพิมพ์ จากผลการทดลองสามารถสรุป อภิปรายผลการดำเนินการวิจัยและข้อเสนอแนะ โดยมีรายละเอียดดังต่อไปนี้

5.1 สรุปและอภิปรายผล

จากผลการประเมินด้วยค่าเฉลี่ย NDCG พบว่า Hybrid2 มีการเรียงลำดับผลลัพธ์การค้นคืนในลำดับที่ 2 ถึง 5 สูงที่สุด ซึ่งเป็นวิธีการนำข้อมูลทางบรรณานุกรมเข้ามาเป็นปัจจัยในการเรียงลำดับใหม่ ได้แก่ จำนวนการอ้างอิงเอกสาร (References) และการถูกอ้างอิงจากเอกสารอื่น (Citation) เป็นตัวแปรที่ใช้วัดคุณภาพของบทความวิจัย จะเห็นว่าเมื่อจำนวนของการอ้างอิงสูง จะส่งผลให้สามารถดันลำดับของบทความขึ้นไปในลำดับต้นๆ ได้ แต่ค่านี้ตัวเลขที่ได้เป็นค่าคงที่ ที่จะไม่มีการเปลี่ยนแปลง ส่วนค่าการถูกอ้างอิงจากบทความวิจัยอื่นๆ โดยค่านี้เป็นตัวแปรที่ส่งผลให้สามารถดันลำดับของเอกสารขึ้นไปอยู่ในลำดับต้นๆ ได้เช่นกัน และเมื่อเวลาผ่านไปจำนวนการถูกอ้างอิงถึงนี้จะต้องเพิ่มขึ้นอย่างต่อเนื่องด้วย เมื่อมีบทความใหม่ๆ เผยแพร่ตีพิมพ์ออกมาใหม่แล้วมี Reference ถึงอย่างต่อเนื่อง ส่วนที่สองคือคุณภาพของแหล่งตีพิมพ์ กำหนดให้บทความวิจัยที่ตีพิมพ์ในวาราสารวิชาการมีค่าน้ำหนักมากกว่างานประชุมวิชาการเนื่องจากกระบวนการคัดกรอง ตรวจสอบจากผู้เชี่ยวชาญ และขั้นตอนในการประเมินที่เน้นคุณภาพมากกว่า และสุดท้ายจำนวนการอ้างอิงถึงมายังบทความวิจัยที่แหล่งตีพิมพ์นี้เป็นผู้จัดพิมพ์ เมื่อระยะเวลาเพิ่มขึ้นแล้วจำนวนการอ้างอิงถึงบทความวิจัยของแหล่งตีพิมพ์ดังกล่าวเพิ่มขึ้นอย่างต่อเนื่อง ซึ่ง

สอดคล้องกับการวัดคุณภาพของบทความวิจัยในส่วนแรก ส่วนของการให้ค่าน้ำหนักข้อมูลแต่ละฟิลด์ (Field Boost) เข้ามาเป็นปัจจัยเสริมร่วมกับการทำ Similarity Ranking นั้นพบว่าถ้า Term ที่อยู่ใน Field ที่เพิ่มค่าน้ำหนักมากๆ มีจำนวน Term น้อย จะส่งผลให้ค่าคะแนนของ Similarity Score นั้นสูงตามไปด้วย เนื่องจากมีการนำความยาวของฟิลด์มาคิดร่วมด้วย ตัวอย่างเช่น ชื่อบทความวิจัย เป็นหัวข้อหลักที่มีความสำคัญมากสุด ให้ค่าน้ำหนักเท่ากับ 3 เป็นฟิลด์ที่มี Term น้อย และค่าน้ำหนักสูงสุดจึงส่งผลให้ค่า Similarity Score สูงกว่าแบบไม่ใช่ Field Boost ซึ่งเมื่อนำปัจจัยทั้งหมดร่วมพิจารณาแล้วทำให้ Hybrid2 มีค่า NDCG ของเอกสารในลำดับต้นๆ ของลิสต์ดีกว่า ดัชนีอื่นๆ

ค่าเฉลี่ยความถูกต้อง (Mean Average Precision: MAP) เป็นการวัดค่าแบบไบนารีคือ จริง จะหมายถึงบทความเกี่ยวข้องกับคำค้น และเท็จ ซึ่งจะหมายถึงบทความที่ค้นคืนออกมาไม่เกี่ยวข้องกับคำค้น เมื่อพิจารณาดูแล้วการประเมินเป็นแบบคะแนน 5 ช่วง คือ 4 – 0 จึงตัดค่าของความถูกต้องที่ 3 โดยช่วงคะแนนที่ 3-4 และ 0-2 พบว่า Hybrid2 ให้ค่าเฉลี่ยความถูกต้องสูงสุด ซึ่งสอดคล้องกับการหาค่าเฉลี่ย NDCG สูง แสดงว่าบทความที่แสดงอยู่ในลำดับต้นๆ นั้นตรงกับความต้องการของผู้ใช้และค่าเฉลี่ยความถูกต้องสูงตามไปด้วยเช่นกัน

การประเมินผลพิจารณาเฉพาะคุณภาพของบทความวิจัยและแหล่งตีพิมพ์ เมื่อบทความมีการเผยแพร่ไปได้ช่วงเวลาหนึ่งจะทำให้ปริมาณการอ้างอิงถึงมีเพิ่มมากขึ้น ส่งผลลำดับการค้นคืนเปลี่ยนไปด้วย ตรงส่วนนี้ทำให้การค้นคืนไม่สามารถค้นคืนบทความวิจัยที่เพิ่งถูกตีพิมพ์ออกมาใหม่ขึ้นมาอยู่ในลำดับต้นๆ ได้ โดยเฉพาะอย่างยิ่งบทความที่เกี่ยวข้องกับเทคโนโลยีใหม่ การคิดค้นประดิษฐ์วิธีการใหม่ๆ ที่ยังไม่เคยมีใครนำมาใช้

5.2 ปัญหาและอุปสรรค

เนื่องจากการวิจัยครั้งนี้ไม่ได้พัฒนาบนระบบค้นคืนบทความวิจัยที่มีข้อมูลบรรณานุกรมและแหล่งตีพิมพ์ เช่น IEEE หรือ ACM Digital Library หรือหากมีการนำต้นแบบนี้ไปใช้งานจริงในระบบค้นคืนบทความวิจัย ซึ่งจำเป็นต้องรวบรวมข้อมูลมาจากเว็บไซต์เดียว ทำให้ต้องใช้ระยะเวลาในการในการ Crawl ใช้เวลานาน ทำให้ต้องหน่วงเวลาในการเข้าถึงเว็บเพจแต่ละครั้ง

5.3 ข้อเสนอแนะ

5.3.1 เพิ่มจำนวนผู้ร่วมทดสอบให้มากขึ้น เพื่อให้ได้ผลการประเมินน่าเชื่อถือมากยิ่งขึ้น

5.3.2 ควรนำปัจจัยทางด้านเวลาเข้ามาเป็นปัจจัยในการสร้างตัวแบบเพิ่มขึ้น เพื่อให้ได้บทความวิจัยที่มีคุณภาพและมีความทันสมัยสามารถถูกค้นคืนมาอยู่ในลำดับต้นๆ ได้

5.3.3 เพิ่มบทความวิจัยในสาขาต่างๆ เพื่อให้ได้ผลการทดลองมีความหลากหลายมากยิ่งขึ้น

บรรณานุกรม

บรรณานุกรม

ภาษาต่างประเทศ

BOOKS

Croft, C., Metzler, D., & Strohman, T. (2009). *Search engines: Information retrieval in practice: International edition*. United States of America: Pearson.

McCandless, M., & Hatcher, E., & Gospodnetic, O. (2010). *Lucene in action, Second edition: Covers Apache Lucene 3.0*. United States of America: Manning.

ARTICLES

Bogers, T., & Bosch, A. V. D. (2008). Recommending scientific articles using CiteULike. In *Proceedings of the 2008 ACM conference on recommender systems* (pp. 287-290). New York, NY: ACM.

Choochaiwattana, W. (2010). Usage of tagging for research paper recommendation. In *International conference on advanced computer theory and engineering* (pp. 439-442). Chengdu, China: IEEE.

Choochaiwattana, W. & Spring, M. B. (2009). Applying social annotation to retrieve and re-rank web resources. In *International conference on information management and engineering* (pp. 215-219). Kuala Lumpur, Malaysia: IEEE.

Geng, X., Liu, T. Y., & Li, H. (2007). Feature selection for ranking. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 407-414). New York, NY: ACM.

Jarvelin, K., & Kekalainen, J. (2000). IR evaluation methods for retrieving highly relevant documents. In *Proceedings of the 23rd annual international ACM SIGIR conference on research and development in information retrieval* (pp. 41-48). New York, NY: ACM.

Jomsri, P. (2011). A combination of similarity ranking and time for social research paper searching. *International Journal of World academy of science, engineering and technology*, 5(6), 574-579. Amsterdam, Netherlands: WASET.

Jomsri, P., Sanguansintukul, S., & Choochaiwattana, W. (2011). CiteRank: Combination similarity and static ranking with research paper searching. *International Journal of Internet technology and secured transactions*, 3(2), 161-177. Geneva, Switzerland: Inderscience.

Lee, D. H., & Brusilovsky, P. (2010). Social networks and interest similarity: The case of CiteULike. In *Proceedings of the 21st ACM conference on hypertext and hypermedia* (pp. 151-156). New York, NY: ACM.

Noel, S., & Beale, R. (2008). Sharing vocabularies: Tag usage in CiteULike. In *Proceedings of the 22nd British HCI group annual conference on people and computers: Culture, Creativity, Interaction* (pp. 71-74). Swinton, UK: British Computer Society.

Parra, D., & Brusilovsky, P. (2009). Collaborative filtering for social tagging system: An experiment with CiteULike. In *Proceedings of the 3rd ACM conference on Recommender systems* (pp. 237-240). New York, NY: ACM.

Pera, M. S., & Ng, Y. K. (2011). A personalized recommendation system on scholarly publications. In *Proceedings of the 20th ACM international conference on information and knowledge management* (pp. 2133-2136). New York, NY: ACM.

Seki, K., Qin, K., & Uehara, K. (2010). Impact and prospect of social bookmarks for bibliographic information retrieval. In *Proceedings of the 10th annual joint conference on Digital libraries* (pp. 357-360). New York, NY: ACM.

Wang, H., He, X., Chang, M., Song, Y., White, R. W., & Chu, W. (2013). Personalized ranking model adaptation for web search. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval* (pp. 323-332). New York, NY: ACM.

Zhuang, Z., & Cucerzan, S. (2006). Re-Ranking search results using query logs. In *Proceedings of the 15th ACM international conference on information and knowledge management* (pp. 860-861). New York, NY: ACM.

ELECTRONIC SOURCES

Class Similarity. Retrieved July 22, 2014, from http://lucene.apache.org/core/3_0_3/api/all/org/apache/lucene/search/Similarity.html

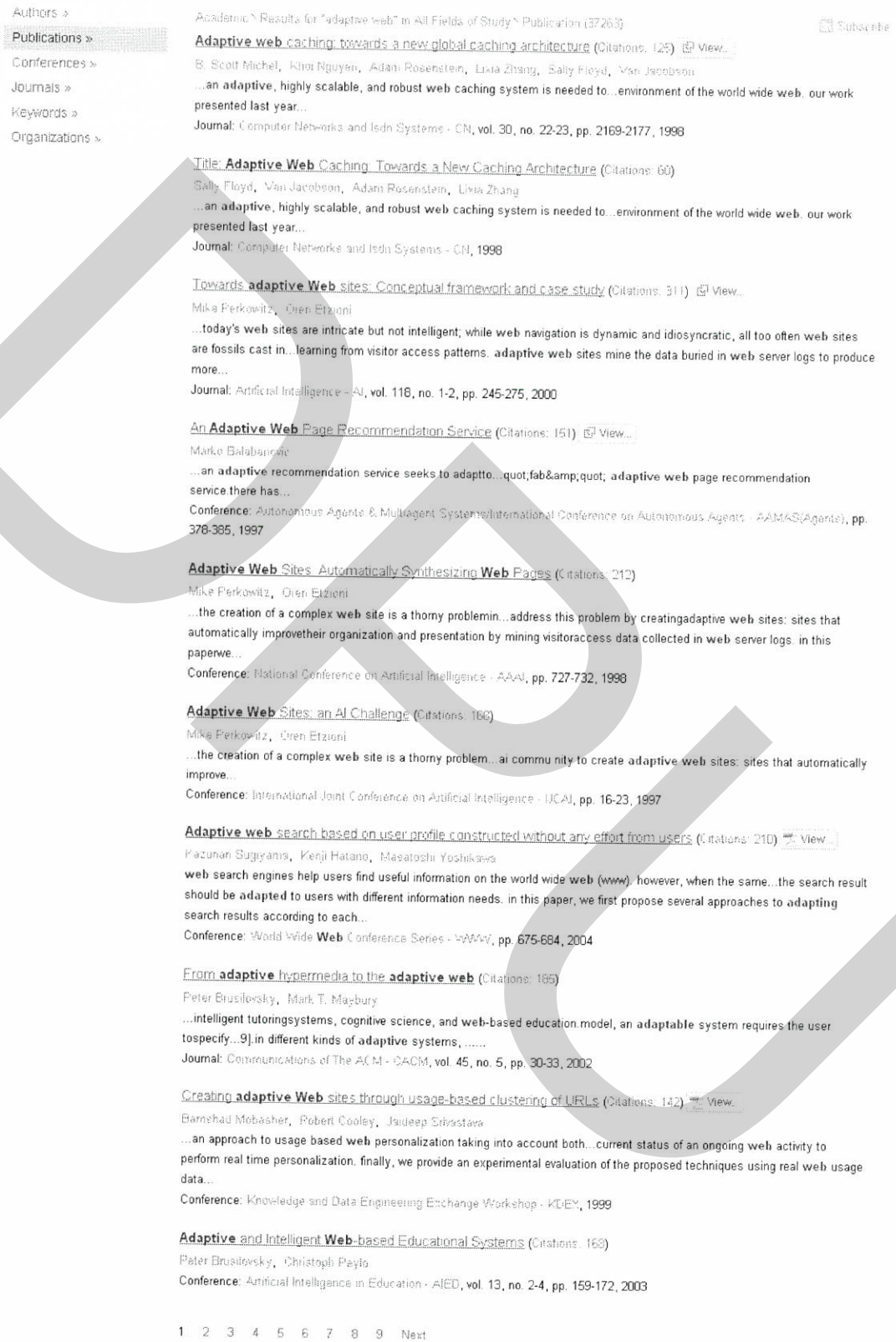
Desktop Search Engine Market Share. Retrieved May 1, 2014, from http://www.netmarketshare.com/search-engine-market-share.aspx?qprid=4&qpcustomd=0

Mean Average Precision. Retrieved July 22, 2014, from https://www.kaggle.com/wiki/MeanAveragePrecision

Search Engine Definition. Retrieved May 1, 2014, from http://searchsoa.techtarget.com/definition/crawler

ภาคผนวก

ภาคผนวก ก
ตัวอย่างการเตรียมคลังเอกสาร



Authors »
Publications »
Conferences »
Journals »
Keywords »
Organizations »

Academic\ Results for "adaptive web" in All Fields of Study\ Publication (37263) Subscribe

Adaptive web caching: towards a new global caching architecture (Citations: 125) View...
B. Scott Michel, Khoi Nguyen, Adam Rosenstein, Lixia Zhang, Sally Floyd, Van Jacobson
...an **adaptive**, highly scalable, and robust **web** caching system is needed to...environment of the world wide **web**. our work presented last year...
Journal: Computer Networks and Isdn Systems - CN, vol. 30, no. 22-23, pp. 2169-2177, 1998

Title: **Adaptive Web** Caching: Towards a New Caching Architecture (Citations: 60)
Sally Floyd, Van Jacobson, Adam Rosenstein, Lixia Zhang
...an **adaptive**, highly scalable, and robust **web** caching system is needed to...environment of the world wide **web**. our work presented last year...
Journal: Computer Networks and Isdn Systems - CN, 1998

Towards **adaptive Web** sites: Conceptual framework and case study (Citations: 311) View...
Mike Perkowitz, Oren Etzioni
...today's **web** sites are intricate but not intelligent; while **web** navigation is dynamic and idiosyncratic, all too often **web** sites are fossils cast in...learning from visitor access patterns. **adaptive web** sites mine the data buried in **web** server logs to produce more...
Journal: Artificial Intelligence - AI, vol. 118, no. 1-2, pp. 245-275, 2000

An **Adaptive Web** Page Recommendation Service (Citations: 151) View...
Marko Balabanovic
...an **adaptive** recommendation service seeks to adaptto...quot;fab" **adaptive web** page recommendation service.there has...
Conference: Autonomous Agents & Multiagent Systems/International Conference on Autonomous Agents - AAMAS(Agents), pp. 378-385, 1997

Adaptive Web Sites: Automatically Synthesizing **Web** Pages (Citations: 212)
Mike Perkowitz, Oren Etzioni
...the creation of a complex **web** site is a thorny problemin...address this problem by creatingadaptive **web** sites: sites that automatically improvetheir organization and presentation by mining visitoraccess data collected in **web** server logs. in this paperwe...
Conference: National Conference on Artificial Intelligence - AAAI, pp. 727-732, 1998

Adaptive Web Sites: an AI Challenge (Citations: 166)
Mike Perkowitz, Oren Etzioni
...the creation of a complex **web** site is a thorny problem...ai commu nity to create **adaptive web** sites: sites that automatically improve...
Conference: International Joint Conference on Artificial Intelligence - IJCAI, pp. 16-23, 1997

Adaptive web search based on user profile constructed without any effort from users (Citations: 210) View...
Kazunari Sugiyama, Kenji Hatano, Masatoshi Yoshikawa
web search engines help users find useful information on the world wide **web** (www). however, when the same...the search result should be **adapted** to users with different information needs. in this paper, we first propose several approaches to **adapting** search results according to each...
Conference: World Wide **Web** Conference Series - WWW, pp. 675-684, 2004

From **adaptive** hypermedia to the **adaptive web** (Citations: 185)
Peter Brusilovsky, Mark T. Maybury
...intelligent tutoringsystems, cognitive science, and **web**-based education.model, an **adaptable** system requires the user tospecify...9].in different kinds of **adaptive** systems, ...
Journal: Communications of The ACM - CACM, vol. 45, no. 5, pp. 30-33, 2002

Creating **adaptive Web** sites through usage-based clustering of URLs (Citations: 142) View...
Bamshad Mobasher, Robert Cooley, Jaideep Srivastava
...an approach to usage based **web** personalization taking into account both...current status of an ongoing **web** activity to perform real time personalization. finally, we provide an experimental evaluation of the proposed techniques using real **web** usage data...
Conference: Knowledge and Data Engineering Exchange Workshop - KDEX, 1999

Adaptive and Intelligent **Web**-based Educational Systems (Citations: 168)
Peter Brusilovsky, Christoph Peylo
Conference: Artificial Intelligence in Education - AIED, vol. 13, no. 2-4, pp. 159-172, 2003

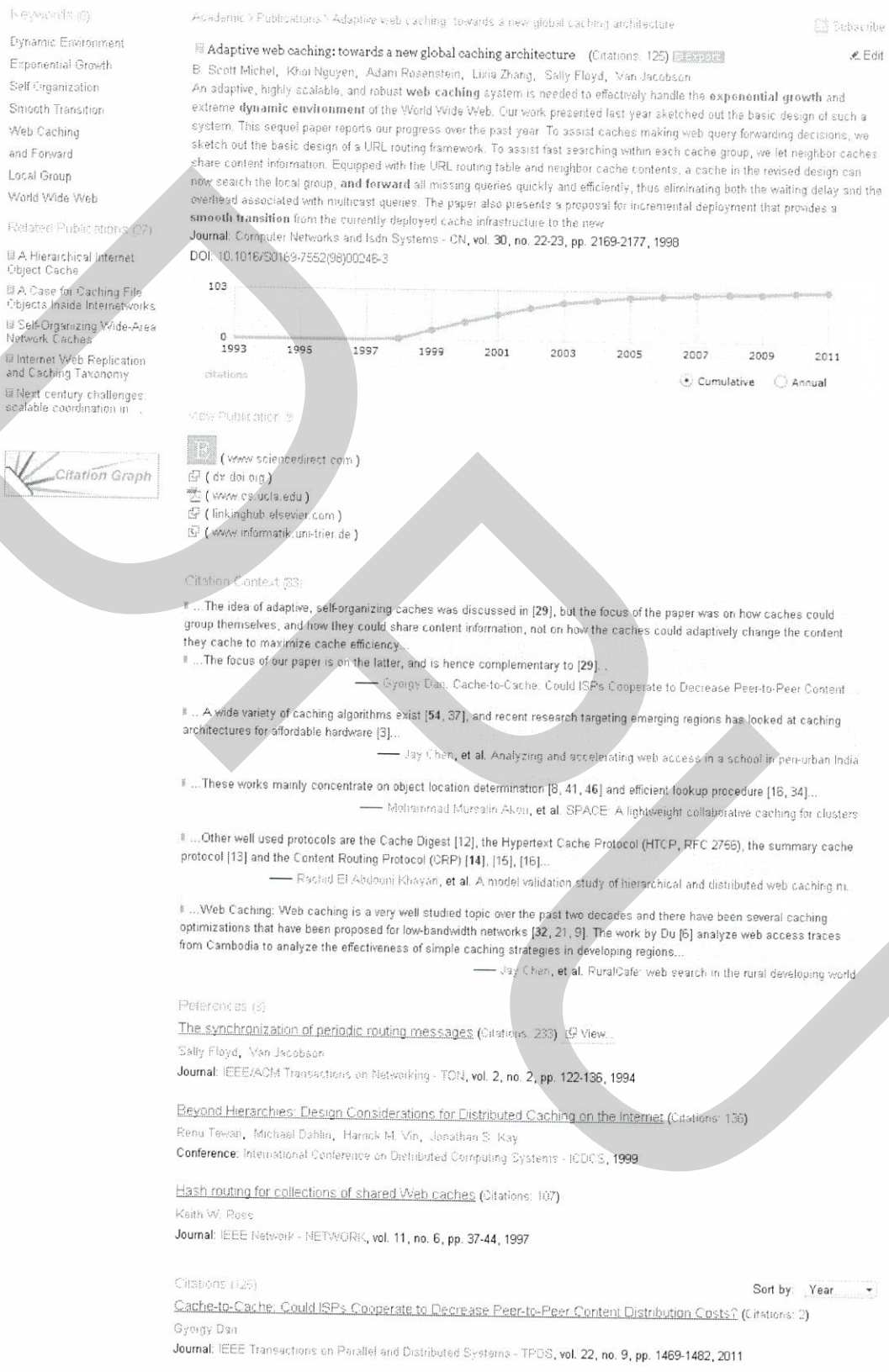
1 2 3 4 5 6 7 8 9 Next

ภาพที่ ก.1 ตัวอย่างหน้ารายการบทความวิจัย

```
▼<li class="paper-item">
  ▼<div class="title-download">
    ▼<div id="ctl00_MainContent_ObjectList_ctl09_divTitle" class="title-fullwidth">
      ▼<h3>
        ▼<a id="ctl00_MainContent_ObjectList_ctl09_Title" onmousedown="try{return
        si(7,'ctl00_MainContent_ObjectList','10');}catch(ex){;}" href="Publication/2659551/adaptive-and-
        intelligent-web-based-educational-systems">
            <b>Adaptive</b>
            " and Intelligent "
            <b>Web</b>
            "-based Educational Systems"
          </a>
        ▶<span id="ctl00_MainContent_ObjectList_ctl09_LbCitation" class="citation">…</span>
        </h3>
      </div>
      ::after
    </div>
  ▶<div class="content">…</div>
    <div class="clear">
        </div>
    <div class="abstract">
            </div>
  ▶<div class="conference">…</div>
    <div class="conference">

            </div>
    <div class="conference">
            </div>
  </li>
</ul>
<div class="clear"></div>
▼<div id="ctl00_MainContent_ObjectList_PageNavigator" class="page-navigator">
  <span class="current">1</span>
  <a href="/Detail?query=adaptive%20web&searchtype=2&start=11&end=20">2</a>
  <a href="/Detail?query=adaptive%20web&searchtype=2&start=21&end=30">3</a>
  <a href="/Detail?query=adaptive%20web&searchtype=2&start=31&end=40">4</a>
  <a href="/Detail?query=adaptive%20web&searchtype=2&start=41&end=50">5</a>
  <a href="/Detail?query=adaptive%20web&searchtype=2&start=51&end=60">6</a>
  <a href="/Detail?query=adaptive%20web&searchtype=2&start=61&end=70">7</a>
  <a href="/Detail?query=adaptive%20web&searchtype=2&start=71&end=80">8</a>
  <a href="/Detail?query=adaptive%20web&searchtype=2&start=81&end=90">9</a>
  <a id="ctl00_MainContent_ObjectList_Next" title="Go to Next Page" class="nextprev" href="/Detail?
  query=adaptive%20web&searchtype=2&start=11&end=20">Next</a>
```

ภาพที่ ก.2 ตัวอย่างหน้าข้อมูลรายการบทความวิจัย



ภาพที่ ก.3 ตัวอย่างหน้ารายละเอียดบทความวิจัย

```
▼ <div class="paper-card">
  ▶ <h1>…</h1>
  ▼ <div id="ctl00_MainContent_PaperItem_divPaper" class="paper-info">
    ▼ <div>
      ▶ <div class="inline-span-right">…</div>
      ▼ <div class="card-title">
          <img id="ctl00_MainContent_PaperItem_imgTitle" title="Paper" class="publication-icon" src=
          "../../images/publication_type_Paper.png" style="border-width:0px;">
          <span id="ctl00_MainContent_PaperItem_title" class="title-span">Adaptive web caching:
          towards a new global caching architecture</span>
          <span style="width: 10px; display: inline-block;">  </span>
        ▶ <span class="citation">…</span>
        </div>
      </div>
    ▼ <div>
        <a class="author-name-tooltip" onmouseover="ViewAuthorTooltip(this,event)" onmouseout=
        "HideAuthorTooltip(this);" href="http://academic.research.microsoft.com/Author/52657200/b-
        scott-michel">B. Scott Michel</a>
        <span class="span-break">, </span>
        <a class="author-name-tooltip" onmouseover="ViewAuthorTooltip(this,event)" onmouseout=
        "HideAuthorTooltip(this);" href="http://academic.research.microsoft.com/Author/52682645/khoi-
        nguyen">Khoi Nguyen</a>
        <span class="span-break">, </span>
        <a class="author-name-tooltip" onmouseover="ViewAuthorTooltip(this,event)" onmouseout=
        "HideAuthorTooltip(this);" href="http://academic.research.microsoft.com/Author/15352/adam-
        rosenstein">Adam Rosenstein</a>
        <span class="span-break">, </span>
        <a class="author-name-tooltip" onmouseover="ViewAuthorTooltip(this,event)" onmouseout=
        "HideAuthorTooltip(this);" href="http://academic.research.microsoft.com/Author/289843/lixia-
        zhang">Lixia Zhang</a>
        <span class="span-break">, </span>
        <a class="author-name-tooltip" onmouseover="ViewAuthorTooltip(this,event)" onmouseout=
        "HideAuthorTooltip(this);" href="http://academic.research.microsoft.com/Author/28377/sally-
        floyd">Sally Floyd</a>
        <span class="span-break">, </span>
        <a class="author-name-tooltip" onmouseover="ViewAuthorTooltip(this,event)" onmouseout=
        "HideAuthorTooltip(this);" href="http://academic.research.microsoft.com/Author/2390511/van-
        jacobson">Van Jacobson</a>
      </div>
```

ภาพที่ ก.4 ตัวอย่างข้อมูลรายละเอียดบทความวิจัย

```
▼<div class="abstract">
 ▼<span id="ctl00_MainContent_PaperItem_snippet">
    "An adaptive, highly scalable, and robust "
    <a href="http://academic.research.microsoft.com/Keyword/44962/web-caching">web caching</a>
    "  system is needed to effectively handle the "
    <a href="http://academic.research.microsoft.com/Keyword/13398/exponential-
    growth">exponential growth</a>
    "  and extreme "
    <a href="http://academic.research.microsoft.com/Keyword/11182/dynamic-environment">dynamic
    environment</a>
    "  of the World Wide Web. Our work presented last year sketched out the basic design of
    such a system. This sequel paper reports our progress over the past year. To assist caches
    making web query forwarding decisions, we sketch out the basic design of a URL routing
    framework. To assist fast searching within each cache group, we let neighbor caches share
    content information. Equipped with the URL routing table and neighbor cache contents, a
    cache in the revised design can now search the local group, "
    <a href="http://academic.research.microsoft.com/Keyword/46357/and-forward">and forward</a>
    "  all missing queries quickly and efficiently, thus eliminating both the waiting delay and
    the overhead associated with multicast queries. The paper also presents a proposal for
    incremental deployment that provides a "
    <a href="http://academic.research.microsoft.com/Keyword/38227/smooth-transition">smooth
    transition</a>
    "  from the currently deployed cache infrastructure to the new"
  </span>
 </div>
 <div style="clear: both;">
      </div>
<div class="conference">

          </div>
▼<div class="conference">
  <span id="ctl00_MainContent_PaperItem_txtJournal">Journal: </span>
  <a id="ctl00_MainContent_PaperItem_HLJournal" class="conference-name" href="../../Journal/
  182/cn-computer-networks-and-isdn-systems">Computer Networks and Isdn Systems - CN</a>
  <span id="ctl00_MainContent_PaperItem_YearJournal" class="year">, vol. 30, no. 22-23, pp.
  2169-2177, 1998</span>
 </div>
 <div class="conference">
          </div>
<div id="divTip1" style="background-color: Silver; position: absolute; display: none;
          padding: 5px; width: 100%;">
       </div>
▶<div id="ctl00_MainContent_PaperItem_divDOI" class="divDOI">…</div>
```

ภาพที่ ก.4 (ต่อ)

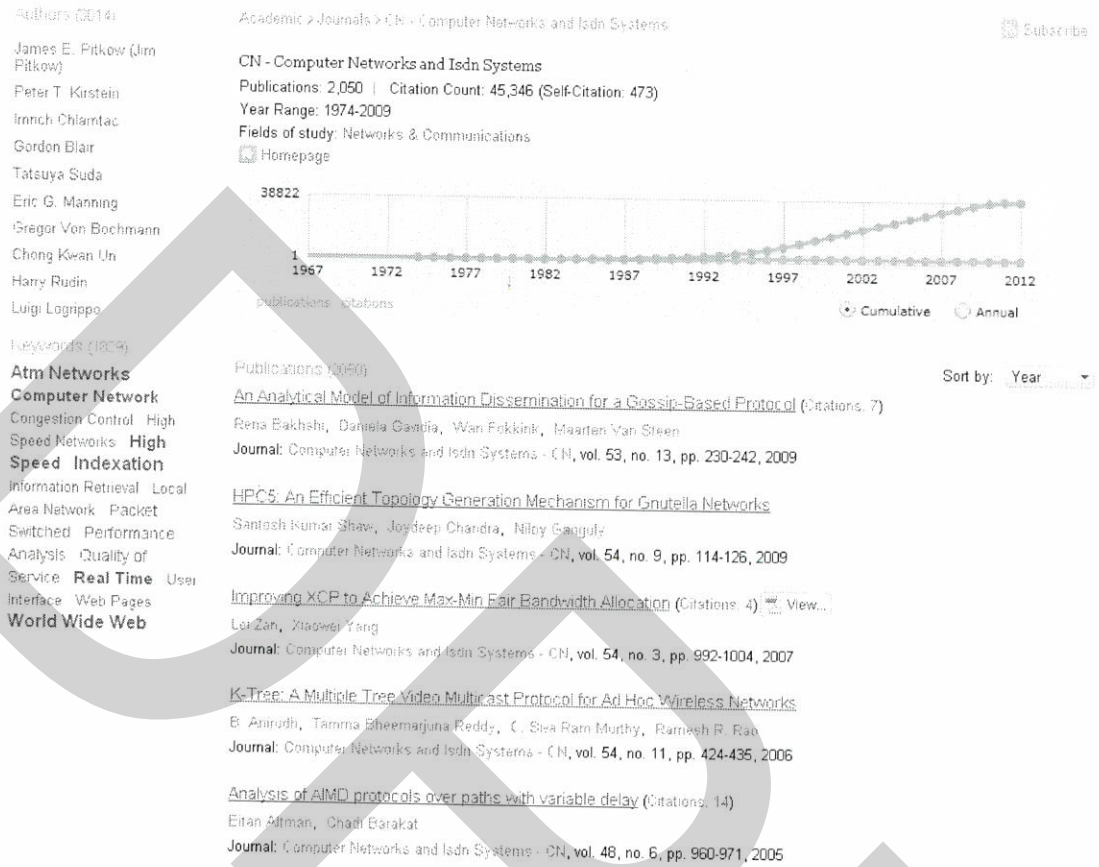
```
▼<div style="float: left">
   <span>DOI:</span>
   <a id="ctl00_MainContent_PaperItem_hypDOIText" href="http://dx.doi.org/10.1016%2fS0169-
   7552(98)00246-3" target="_blank">10.1016/S0169-7552(98)00246-3</a>
  </div>
  <div style="float: left;">

            </div>
 </div>
</div>
<div style="clear: both;">
   </div>
▶<div id="ctl00_MainContent_PaperItem_divTrendChartDownLoad">…</div>
</div>
▼<div class="section-wrapper">
 ▼<div id="ctl00_MainContent_PaperCitationContextList_ListHeader" class="section-header">
  ▼<div class="title">
   ▼<h2>
    ▼<a id="ctl00_MainContent_PaperCitationContextList_ctl00_HeaderLink" title="Show the context
    of how this paper is referenced by others" href="../../Detail?
    entitytype=1&searchtype=7&id=1269614">
       "Citation Context "
       <span class="item-count">(83)</span>
     </a>
    </h2>
   </div>
  <div id="ctl00_MainContent_PaperCitationContextList_Catalog" class="list-tab">


     </div>
 </div>
▼<div class="section-wrapper">
 ▼<div id="ctl00_MainContent_PaperList_ListHeader" class="section-header">
  ▼<div class="title">
   ▼<h2>
    ▼<a id="ctl00_MainContent_PaperList_ctl00_HeaderLink" href="../../Detail?
    entitytype=1&searchtype=2&id=1269614">
       "References "
       <span class="item-count">(3)</span>
     </a>
    </h2>
   </div>
  <div id="ctl00_MainContent_PaperList_Catalog" class="list-tab">


     </div>
 </div>
▶<ul>…</ul>
 <div class="clear"></div>
</div>
▼<div class="section-wrapper">
 ▼<div id="ctl00_MainContent_CitationList_ListHeader" class="section-header">
  ▶<div class="orderby-filter">…</div>
  ▶<script type="text/javascript" language="javascript">…</script>
  ▼<div class="title">
   ▼<h2>
    ▼<a id="ctl00_MainContent_CitationList_ctl00_HeaderLink" href="../../Detail?
    entitytype=1&searchtype=5&id=1269614">
       "Citations "
       <span class="item-count">(125)</span>
     </a>
    </h2>
   </div>
  <div id="ctl00_MainContent_CitationList_Catalog" class="list-tab">
```

ภาพที่ ก.4 (ต่อ)



ภาพที่ ก.5 ตัวอย่างหน้ารายละเอียดแหล่งตีพิมพ์ และการเผยแพร่บทความวิจัย

```
▼<div class="conference-card">
   <h1>Computer Networks and Isdn Systems,CN,Networks & Communications</h1>
  ▼<div class="content">
   ▼<div class="card-title">
      <span id="ctl00_MainContent_JournalItem_name">CN - Computer Networks and Isdn Systems</span>
    </div>
   </div>
  ▼<div class="detailInfo">
   ▼<span id="ctl00_MainContent_JournalItem_detailInfo">
      "Publications: 2,050"
      <span class="space">|</span>
      <span>Citation Count: 45,346 (Self-Citation: 473)</span>
      <br>
      "Year Range: 1974-2009"
    </span>
   </div>
  ▼<div>
     "
           Fields of study: "
    <a href="http://academic.research.microsoft.com/RankList?
    entitytype=4&topDomainID=2&subDomainID=14&last=0&start=1&end=100">Networks & Communications</a>
   </div>
  ▶<div id="ctl00_MainContent_JournalItem_HomePageDiv" class="homepage">…</div>
  ▶<div id="holder" class="trendchart">…</div>
  ▶<div class="trendchart-paperlist" id="divPaperList">…</div>
  ▶<script language="javascript" type="text/javascript">…</script>
 </div>
▶<div class="section-wrapper">…</div>
</div>
```

ภาพที่ ก.6 ตัวอย่างข้อมูลรายละเอียดแหล่งตีพิมพ์ และการเผยแพร่บทความวิจัย

ภาคผนวก ข

การออกแบบตารางฐานข้อมูล

ตารางที่ ข.1 ตาราง Article เก็บรายละเอียดข้อมูลทางบรรณานุกรมของบทความวิจัย

ลำดับ	แอทริบิวต์	ความหมาย	ชนิด	คีย์
1	articleId	รหัสบทความวิจัย	INTEGER	PK
2	articleUrl	ลิงค์บทความวิจัย	TEXT	
3	title	หัวข้อบทความวิจัย	TEXT	
4	abstracts	บทคัดย่อ	TEXT	
5	author	ผู้แต่ง	TEXT	
6	doi	Digital Object Identifier	TEXT	
7	issn	International Standard Serial Number	TEXT	
8	number	หมายเลขบทความวิจัย	TEXT	
9	pages	เลขที่หน้าบทความวิจัย	TEXT	
10	volume	ปีพิมพ์	TEXT	
11	year	ค.ศ. ปีพิมพ์	TEXT	
12	citationContext	จำนวนงานวิจัยที่อ้างอิงถึงและปรากฏอยู่ในเนื้อหา	INTEGER	
13	reference	จำนวนเอกสารอ้างอิง	INTEGER	
14	citationList	จำนวนงานวิจัยที่อ้างอิงถึง	INTEGER	
15	iPage	ดัชนีลำดับการเข้าถึงเนื้อหาหน้าเว็บ	INTEGER	
16	status	สถานะการอ่านข้อมูลจาก URL	INTEGER	
17	publishId	รหัสแหล่งตีพิมพ์	INTEGER	FK

ตารางที่ ข.2 ตาราง Tag เก็บรายละเอียดคำสำคัญ

ลำดับ	แอทริบิวต์	ความหมาย	ชนิด	คีย์
1	tagId	รหัสคำสำคัญ	INTEGER	PK
2	tagName	คำสำคัญ	TEXT	

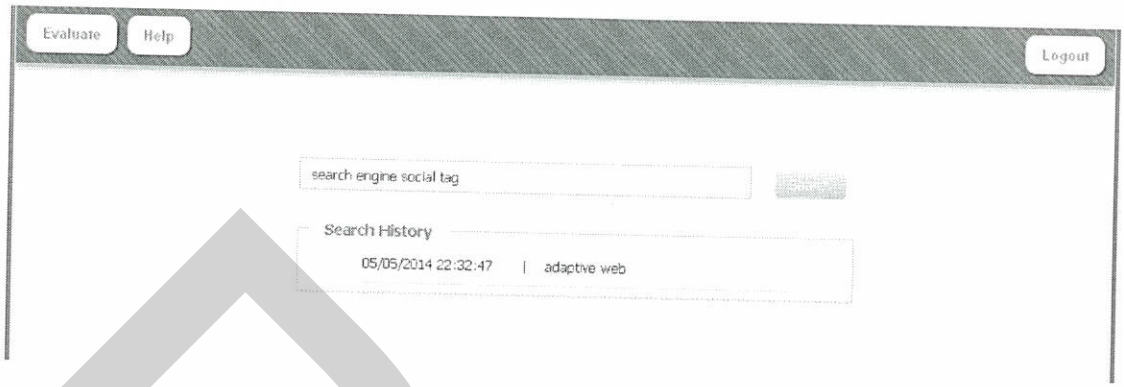
ตารางที่ ข.3 ตาราง Publisher เก็บรายละเอียดสำนักพิมพ์

ลำดับ	แอทริบิวต์	ความหมาย	ชนิด	คีย์
1	publishId	รหัสแหล่งตีพิมพ์	INTEGER	PK
2	publishUrl	URL ที่เข้าถึงข้อมูลแหล่งตีพิมพ์	TEXT	
3	publishTitle	ชื่อแหล่งตีพิมพ์ที่ปรากฏบนเว็บ	TEXT	
4	publishName	ชื่อแหล่งตีพิมพ์	TEXT	
5	publishType	ประเภทการตีพิมพ์ 0 Journal 1 Conference	INTEGER	
6	publications	จำนวนบทความวิจัยที่ถูกตีพิมพ์	INTEGER	
7	citationCount	จำนวนบทความวิจัยที่อ้างอิงถึง	INTEGER	
8	selfCitation	จำนวนงานวิจัยที่อ้างอิงถึงจาก ภายในสำนักพิมพ์เดียวกัน	INTEGER	
9	yearRange	ช่วงปีที่มีการตีพิมพ์	TEXT	
10	startYear	ปีที่เริ่มตีพิมพ์	INTEGER	
11	endYear	ปีล่าสุดที่มีการตีพิมพ์	INTEGER	
12	fieldsOfStudy	ประเภทการวิจัย	TEXT	
13	status	สถานะการอ่านข้อมูลจาก URL	INTEGER	

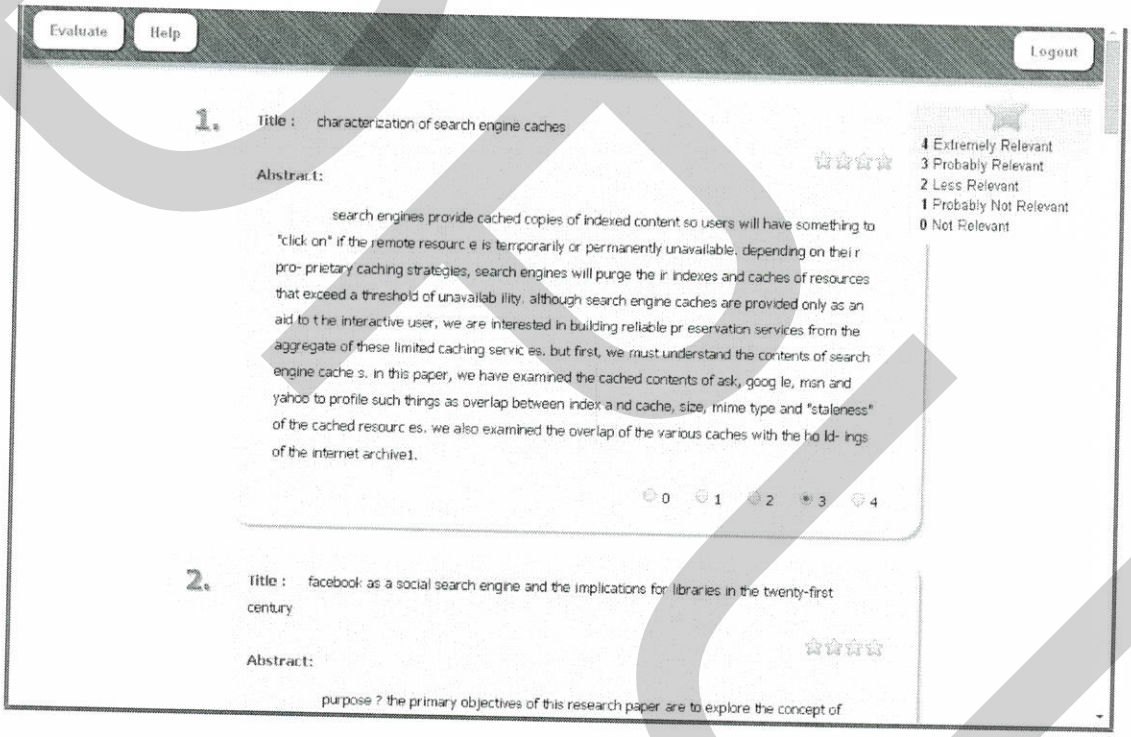
ตารางที่ ข.4 ตาราง AriticleTag เก็บรายละเอียดความสัมพันธ์ของบทความวิจัยกับจำนวนคำสำคัญ

ลำดับ	แอทริบิวต์	ความหมาย	ชนิด	คีย์
1	articleId	รหัสบทความวิจัย	INTEGER	PK
2	tagId	รหัสคำสำคัญ	INTEGER	PK
3	count	จำนวนคำสำคัญที่ปรากฏอยู่ใน บทคัดย่อ	INTEGER	

ภาคผนวก ค
ตัวอย่างหน้าจอระบบค้นคืนบทความวิจัย



ภาพที่ ค.1 หน้าจอสำหรับสืบค้นบทความวิจัย



ภาพที่ ค.2 ตัวอย่างหน้าทำประเมินผลลัพธ์การค้นคืน

ภาคผนวก ง
ตัวอย่างผลการประเมินจากผู้ทดสอบ

ตารางที่ ง.1 ตัวอย่างการประเมิน Judgment Score

Title	Abstract	SimScore	JudgeScore
oiled: a reasonable ontology editor for the semantic web	ontologies will play a pivotal role in the semantic web, where they will provide a source of precisely dened terms that can be communicated across people and applications. oiled, is an ontology editor that has an easy to use frame interface, yet at the same time allows users to exploit the full power of an expressive web ontology language (oil). oiled uses reasoning to support ontology design, facilitating the development of on- tologies that are both more detailed and more accurate.	1.000000	2
ontology versioning on the semantic web	ontologies are often seen as basic building blocks for the semantic web, as they provide a reusable piece of knowledge about a specific domain. however, those pieces of knowledge are not static, but evolve over time. domain changes, adaptations to different tasks, or changes in the conceptualization require modifications of the on- tology. the evolution of ontologies causes operability problems, which will hamper their effective reuse. a versioning mechanism might help to reduce those problems, as it will make the relations between different revisions of an ontology explicit. this paper will discuss the problem of ontology versioning, inspired by the work done in database schema versioning and program interface versioning, it will also propose building blocks for the most important aspects of a versioning mechanism, i.e., ontol- ogy identification and change specification.	0.967174	3
domain ontology component-based semantic information integration	research on architecture of domain ontology component-based information semantic representation and integration is studied. domain ontology component, a "loosely coupled" approach in the use of ontology, is advocated. as a case study, a prototype for agricultural policy-oriented domain ontology component-based semantic information integration system (apodocsiis) is established. ontology plays a key role in providing a shared terminology and supporting for the semantic representation and integration process. the architecture allows apodocsiis-based applications to perform automatic semantic information integration of agricultural policy text at more length: semantic matching of concepts between different ontology components, domain ontology component-based dynamic semantic annotation of unstructured and semi-structured content, semantically-enabled information extraction, indexing, retrieval, integration, as well as ontology management, such as querying and modifying the underlying ontology components. main frame of this architecture have been implemented and concrete integration example are given.	0.921588	3

ตารางที่ ง.1 (ต่อ)

Title	Abstract	SimScore	JudgeScore
alysis of the origin of ontology mismatches on the semantic web	despite the potential of domain ontologies to provide consensual representations of domain-relevant knowledge, the open, distributed and decentralized nature of the semantic web means that individuals will rarely, if ever, countenance a common set of terminological and representational commitments during the ontology design process. more often than not, differences between ontologies are likely to occur, and this is the case even when the ontologies describe identical or overlapping domains of interest. differences between ontologies are often referred to as ontology mismatches and there is an extensive research literature geared towards the technology- mediated reconciliation of such mismatches. our approach in the current paper is not to comment on the relative merits or demerits of the various technological solutions that could be used to resolve ontological differences; rather, we aim to explore the reasons why such differences may arise in the first place. in addition to a review of the various factors that contribute to ontology mismatches on the semantic web, we also discuss a number of focus areas for future research in this area. an improved understanding of the origins of ontology mismatches will, we argue, complement existing research into semantic integration techniques. in particular, by understanding more about the complex cognitive, epistemic and socio-cultural factors associated with the ontology development process, we may be able to develop knowledge acquisition and modeling tools/techniques that attenuate the impact of ontology mismatches for large-scale information sharing and data integration on the semantic web.	0.896974	3
semantic annotation of data tables using a domain ontology	in this paper, we show the different steps of an annotation process that allows one to annotate data tables with the relations of a domain ontology. the columns of a table are first segregated according to whether they represent numeric or symbolic data. then, we annotate the numeric columns with their corresponding numeric type, and the symbolic columns with their corresponding symbolic type, combining different evidences from the ontology. the relations represented by a table are recognized using both the table title and the types of the columns. we give experimental results for our annotation method.	0.942749	2

ภาคผนวก จ

บทความการประชุมวิชาการระดับชาติ ด้านคอมพิวเตอร์และเทคโนโลยีสารสนเทศ (NCCIT)
ครั้งที่ 10 คณะเทคโนโลยีสารสนเทศ มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ

ตัวแบบสำหรับการเรียงลำดับผลลัพธ์การค้นคืนในระบบค้นคืนบทความวิจัยโดยการใช้ข้อมูลทางบรรณานุกรม

A Model for Ranking Search Results in a Research Paper Search Engine Using Bibliographic Information

ขวัญเรือน โสอุบล (Khwanruan So-Ubol) และ วรสิทธิ์ ชูชัยวัฒนา (Worasit Choochaiwattana)
สาขาวิชาวิศวกรรมเว็บ คณะเทคโนโลยีสารสนเทศ มหาวิทยาลัยธุรกิจบัณฑิตย์
ksoubol@gmail.com, worasit.cha@dpu.ac.th

บทคัดย่อ

Query Dependent Ranking หรือ Similarity Ranking เป็นเทคนิคสำหรับเรียงลำดับผลลัพธ์การค้นคืนโดยการเปรียบเทียบคำค้นและดัชนีของเอกสาร ซึ่งไม่ได้พิจารณาถึงปัจจัยอื่นๆ เช่น คุณภาพของเอกสาร ในขณะที่ Query Independent Ranking หรือ Static Ranking เป็นเทคนิคที่สำคัญอีกเทคนิคหนึ่ง สำหรับการเรียงลำดับผลลัพธ์การค้นคืน โดยพิจารณาคุณภาพของเอกสารเป็นหลัก ในงานวิจัยนี้เสนอตัวแบบสำหรับเรียงลำดับผลลัพธ์การค้นคืนในระบบค้นคืนบทความวิจัย ที่มีการผสมผสานระหว่าง Query Dependent Ranking และ Query Independent Ranking โดยจะนำเอาข้อมูลทางบรรณานุกรม มาใช้สำหรับการประเมินคุณภาพของบทความวิจัย ซึ่งจะเป็นการใช้ Similarity Feature ประกอบด้วย ชื่องานวิจัย บทคัดย่อ คำสำคัญ มาประยุกต์ร่วมกับข้อมูลทางบรรณานุกรมของงานวิจัยแต่ละงานซึ่งประกอบไปด้วย จำนวนการอ้างอิง จำนวนการถูกอ้างอิงถึง และข้อมูลแหล่งการตีพิมพ์งานวิจัย ได้แก่ งานประชุมวิชาการ หรือวารสารวิชาการ จำนวนเอกสารที่มีการตีพิมพ์ในงานประชุมวิชาการหรือวารสารวิชาการ จากผลการทดสอบเบื้องต้น พบว่าการผสมผสานระหว่าง Query Dependent Ranking และ Query Independent Ranking สามารถให้ผลการค้นคืนเอกสาร ที่มีคุณภาพในห้าลำดับแรกดีกว่าวิธีการสืบค้นแบบใช้ Query Dependent Ranking เพียงอย่างเดียว

คำสำคัญ: การเรียงลำดับผลลัพธ์การค้นคืน ตัวแบบ ระบบค้นคืนบทความวิจัย

Abstract

Query Dependent Ranking or Similarity Ranking is a technique for ranking search results by comparing query terms with document indexes. This technique doesn't consider other related factors such as quality of documents. While, Query Independent Ranking or Static Ranking is another important ranking search results technique by focusing on quality of documents. This research paper proposed a model for ranking search results in a research paper search engine. The proposed a technique to combine Query Dependent Ranking with Query Independent Ranking using bibliographic information. The bibliographic information was used to determine a quality of research papers. This technique started with the usage of similarity feature, such as title of research papers, abstract, and keywords, in combination with a bibliographic information of each research paper, such as number of citation, number of cited by other papers, and source of publication including conference proceeding, peer-review journal. From the preliminary result of experiment, the combination technique between Query Dependent Ranking and Query Independent Ranking provide more relevant research paper search results for the top five ranking results comparing with the results from the Query Dependent Ranking technique only.

Keyword: Ranking Search Results, Model, Research Paper Search Engine

1. บทนำ

เทคโนโลยีสารสนเทศและอินเทอร์เน็ตถูกพัฒนา ไปอย่างรวดเร็ว ทำให้ปริมาณข้อมูลและสารสนเทศต่างๆ ถูกเผยแพร่มากมายมหาศาล ดังนั้นการพัฒนาระบบการสืบค้นข้อมูลที่มีประสิทธิภาพ และตรงกับความต้องการของผู้ใช้จึงทำได้ยากมากขึ้นด้วย

ระบบสืบค้นที่นำมาใช้งานในอดีต เช่น การสืบค้นข้อมูลของ Yahoo! ใช้วิธีการที่เรียกว่า Catalog Based Information Retrieval จะเป็นการสืบค้นจากหมวดหมู่หลัก แล้วย่อยลงไปจนถึงหัวข้อที่ต้องการ หรืออีกวิธีการหนึ่งที่นิยมใช้กันอย่างแพร่หลายก็คือ Query Based Search Engine เป็นการสืบค้นข้อมูลที่มักจะพิจารณา โดยเปรียบเทียบความเหมือนระหว่างคำค้น (Query) และดัชนีของเอกสารเท่านั้น เอกสารที่ได้จากผลลัพธ์การค้นคืนจะไม่มีความสัมพันธ์กับคำค้นที่ต้องการเลย ซึ่งวิธีการดังกล่าวถูกเรียกว่า Query Dependent Ranking หรือ Similarity Ranking

ระบบสืบค้นในยุคถัดมา เช่น กูเกิ้ล (Google) เริ่มมีการนำเอาปัจจัยที่เกี่ยวข้อง อย่างคุณภาพของเอกสารเข้ามาร่วมพิจารณาในการเรียงลำดับผลลัพธ์การค้นคืน โดยวิธีการดังกล่าวถูกเรียกว่า Query Independent Ranking หรือ Static Ranking [1]

ในการศึกษางานวิจัยนี้ทำการทดลองเพื่อพิสูจน์สันนิษฐานว่า เทคนิคสำหรับการเรียงลำดับแบบ Query Dependent Ranking แบบผสมผสานข้อมูลบรรณานุกรม ให้ผลลัพธ์ดีกว่า Query Dependent Ranking เพียงอย่างเดียว โดยการสร้างดัชนีต้นแบบ ของทั้งสองวิธี และเพื่อให้ผลการทดลองสามารถควบคุมปัจจัย ภายในและขอบเขตของข้อมูล ในผลงานชิ้นนี้จึงใช้มุ่งเน้นไปที่ข้อมูลบทความวิจัย

ในส่วนที่ 2 กล่าวถึงงานวิจัยที่เกี่ยวข้องกับระบบสืบค้น (Search Engine) และระบบสืบค้นบทความวิจัย (Research Paper Search Engine) ที่ผ่านมา ส่วนที่ 3 การนำเสนอตัวแบบประกอบด้วยภาพรวมของระบบ ตั้งแต่ขั้นตอนการเก็บข้อมูลเพื่อสร้างดัชนีต้นแบบ ใช้ตัวแทนของเอกสารทั้งหมดในระบบสืบค้นที่อยู่ในคลังเอกสาร (Paper Corpus) ส่วนที่ 4 กล่าวถึงวิธีการทดลอง ส่วนที่ 5 การประเมินผลระบบสืบค้นของแต่ละดัชนี และการอภิปรายผลในหัวข้อสุดท้าย

2. งานวิจัยที่เกี่ยวข้อง

ในกระบวนของระบบสืบค้นสารสนเทศ มีวิธีการนำเสนอที่หลากหลาย ขึ้นอยู่กับลักษณะและประเภทของสารสนเทศ ในช่วงเวลา 5 ปีที่ผ่านมา งานวิจัยที่เกี่ยวข้องมักจะใช้ข้อมูลต่างๆ จากระบบเครือข่ายทางสังคม (Social Networking System) เช่น CiteULike และ del.io.us โดยนำ Social Tag และข้อมูลที่เกี่ยวข้องกับการ Bookmark มาใช้เพื่อทำ Query Dependent Ranking และ Query Independent Ranking [2] นอกจากนั้นแล้วยังนำเอาข้อมูลดังกล่าวมาทำ Profile ของผู้ใช้งาน เพื่อสร้างระบบแนะนำเนื้อหา (Recommendation System) [3]. [4]. [5]. [6]. [7]. [8]

สำหรับงานวิจัยที่เกี่ยวข้องกับการเรียงลำดับผลลัพธ์การค้นคืนของบทความวิจัยนั้น มีการเสนอวิธีการในการเรียงลำดับผลลัพธ์โดยการใช้เวลา [9] และใช้ข้อมูลอื่น ๆ เช่น จำนวนของกลุ่มผู้ใช้ที่มีการแชร์บทความ จำนวนผู้ที่ชื่นชอบบทความ และเวลา (Time Stamp) ของการแบ่งปันบทความ [10] ซึ่งจะพบว่ายังมีข้อมูลอื่น ๆ ที่น่าสนใจเอามาใช้ในการประเมินคุณภาพของบทความวิจัย เพื่อทำการจัดลำดับผลลัพธ์การค้นคืนอีก

จากวิจัยที่กล่าวมา พบว่าบทความวิจัยที่มีการนำเสนอไว้ก่อนหน้ามีการนำข้อมูลแวดล้อมของเอกสารมาพิจารณาเพิ่มและถูกนำมาใช้เรียงลำดับผลลัพธ์ โดยมุ่งเน้นไปที่ลักษณะของผู้ใช้งาน (Profile) มากกว่ามุ่งเน้นไปที่คุณภาพของเอกสารที่ได้ ดังนั้นในงานวิจัยนี้จึงเสนอตัวแบบ สำหรับเรียงลำดับผลลัพธ์การค้นคืนในระบบค้นคืนบทความวิจัย ที่มีการผสมผสานระหว่าง Query Dependent Ranking และ Query Independent Ranking โดยจะนำเอาข้อมูลทางบรรณานุกรมมาใช้สำหรับการประเมินคุณภาพของบทความวิจัย โดยจะเน้นที่การประเมินคุณภาพของแหล่งตีพิมพ์ และข้อมูลการอ้างอิงเป็นหลัก

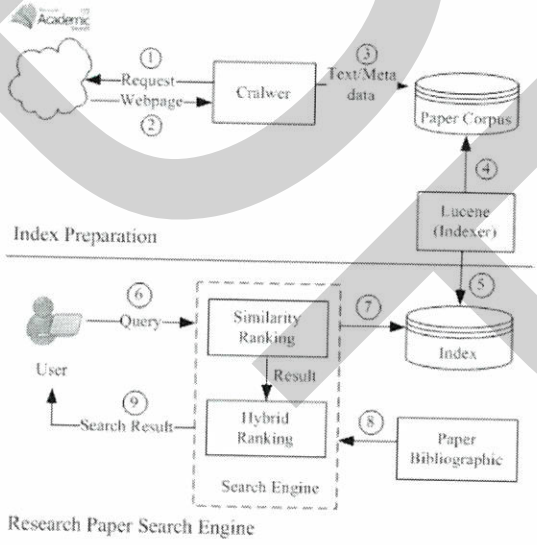
3. การนำเสนอตัวแบบ

การศึกษาวิจัยมีการสร้างระบบสืบค้นขึ้นเพื่อใช้พิสูจน์ ตัวแบบ มีขั้นตอนการทำงาน แสดงดังภาพที่ 1 ดังนี้

3.1 Crawler

เป็นโปรแกรมที่ทำหน้าที่เก็บข้อมูลจากอินเทอร์เน็ต เพื่อวิเคราะห์ และกรองรายละเอียดที่ต้องเก็บลงฐานข้อมูล โดยใน

การศึกษาใช้ข้อมูลจาก academic.research.microsoft.com ซึ่งเป็นผู้ให้บริการข้อมูลบทความวิจัยสาขาต่างๆ ในแต่ละบทความมีรายละเอียดทางบรรณานุกรม เช่น ชื่อหัวข้อบทความวิจัย (Title) ผู้แต่ง (Author) คำสำคัญ (Keyword) บทคัดย่อ (Abstract) ปีที่ตีพิมพ์ (Year) จำนวนการถูกอ้างอิงถึง (Citation) จำนวนเอกสารอ้างอิง (Reference) และรายละเอียดของงานประชุมวิชาการที่บทความได้ตีพิมพ์ (Publisher) เช่น บทความนี้ถูกตีพิมพ์ในงานประชุมวิชาการ (Conference) หรือวารสารวิชาการ (Journal) จำนวนบทความวิจัยที่ถูกตีพิมพ์แล้ว (Publication) ระยะเวลาปีที่มีการตีพิมพ์ (Year Range) และจำนวนเอกสารมีการอ้างอิงถึง (Publisher Citation)



ภาพที่ 1: ขั้นตอนการทำงานของระบบ

3.2 Paper Corpus

เป็นคลังเอกสารที่เก็บรวบรวมและบันทึกบทความวิจัยทั้งหมดของระบบที่ได้จาก Crawler

3.3 ดัชนี (Index)

ดัชนีเป็นขั้นตอนการเตรียมข้อมูล โดยใช้ Lucene เป็นไลบรารี่สำหรับคำนวณค่าความถี่ของคำในเอกสาร (Term) ที่ปรากฏอยู่ในเอกสารทั้งหมด สำหรับการสืบค้นที่มีประสิทธิภาพและรวดเร็ว จัดเก็บอยู่ในรูปแบบของเว็กเตอร์ (Vector) วิธีการนี้จะถูกนำมาใช้ในงาน Information Retrieval และ Text Mining ซึ่งจะเห็นว่างานส่วนใหญ่จะเป็นการเปรียบเทียบเอกสารที่อยู่ใน Corpus กับคำค้นเท่านั้น

คำนวณค่าน้ำหนัก $w_{t,D}$ ของคำ หมายถึงคำแต่ละคำมีความเกี่ยวข้อง (Relevant) กับเอกสารมากน้อยแค่ไหน ได้จาก *tf.idf* แสดงดังสมการที่ (1) และ (2)

$$tf.idf_{t,D} = tf_{t,D} \times idf_t \quad (1)$$

$$w_{t,D} = tf_{t,D} \times log\left(\frac{N}{df_t}\right) \quad (2)$$

เมื่อ $w_{t,D}$ แทนค่าน้ำหนักของ $tf.idf_{t,D}$ และหมายถึงความเกี่ยวข้องของคำ t ในเอกสาร D

$tf_{t,D}$ (Term Frequency) แทนจำนวนคำ t ที่ปรากฏในเอกสาร D

N แทนจำนวนเอกสารทั้งหมด

idf_t (Inverse Document Frequency) แทนจำนวนเอกสารที่มี t ปรากฏอยู่ จากสมการ (2) พบว่าเมื่อ t ปรากฏอยู่ในทุกเอกสาร N ส่งผลให้ค่า idf_t มีค่าความสำคัญลดลงจนมีค่าเป็นศูนย์

3.4 Re-Ranking Model

จากสมการที่กล่าวไปก่อนหน้า ทำการทดลองโดยสร้างต้นแบบดัชนีทั้งหมด 4 แบบ ดังนี้

1. Index0 แทน Full-Text Index

2. Index 1 แทน Full-Text Boost Field Index มีการเพิ่มค่าน้ำหนักให้กับฟิลด์ข้อมูล (Boost Field) ได้แก่ ชื่อ (Title) บทคัดย่อ (Abstract) และ คำสำคัญ (Keyword) ค่าน้ำหนักแต่ละตัวเป็น 3 2 และ 1 ตามลำดับ

3. Hybrid0 แทน Full-Text Index กับ Bibliographic

4. Hybrid1 แทน Full-Text Boost Field Index ได้แก่ ชื่อ (Title) บทคัดย่อ (Abstract) และ คำสำคัญ (Keyword) ค่าน้ำหนักแต่ละตัวเป็น 3 2 และ 1 ตามลำดับ กับ Bibliographic

เมื่อ Index0 และ Index1 ใช้คุณสมบัติจาก Similarity Feature และ Hybrid0 และ Hybrid1 ใช้ Similarity Feature ร่วมกับ Bibliographic Feature สามารถหาค่าน้ำหนักของเอกสารดังนี้

ความสัมพันธ์ระหว่าง Similarity Feature กับ Bibliographic Feature ของเอกสารงานวิจัยและผู้จัดพิมพ์ ตามสมการที่ (3)

$$Hybrid\ Score = Sim(\propto) + Bib(1-\propto) \quad (3)$$

เมื่อ *Hybrid Score* แทนค่าคะแนนที่เกิดจากการวัดคุณภาพของ Similarity Score ร่วมกับ Bibliographic Score

Sim แทน Similarity Score

Bib แทน Bibliographic Score คิดจากค่าเฉลี่ยจากการวัดค่าคุณภาพของบทความวิจัย ร่วมกับคุณภาพของผู้จัดพิมพ์งานวิจัย ได้จากสมการที่ (4) และ (5) ตามลำดับ

$$QA = R(\beta) + CA(1-\beta) \quad (4)$$

QA แทนคุณภาพของเอกสารงานวิจัย (Article Quality) ประกอบด้วย *R* และ *CA* และแทนค่า β เท่ากับ 0.9 ค่าน้ำหนักที่กำหนดให้

R แทนจำนวนเอกสารอ้างอิงภายในบทความวิจัยนั้น ค่าที่นำมาใช้ทำ Scale Normalized อยู่ระหว่าง 0 ถึง 1 คำนวณจากเอกสาร 30 เอกสารแรกที่ได้จากการค้นคืน

CA แทนจำนวนเอกสารที่มีการอ้างอิงถึงบทความวิจัยนั้น ค่าที่นำมาใช้ทำ Scale Normalized อยู่ระหว่าง 0 ถึง 1 คำนวณจากเอกสาร 30 เอกสารแรกที่ได้จากการค้นคืน

$$QP = Type \times CP \quad (5)$$

QP แทนคุณภาพของผู้จัดพิมพ์งานวิจัย (Publisher Quality)

Type แทนประเภทของผู้จัดพิมพ์ กำหนดให้วารสารวิชาการ (Journal) เท่ากับ 1.0 งานประชุมวิชาการ (Conference) เท่ากับ 0.1

CP แทนจำนวนเอกสารที่มีการอ้างอิงถึงผู้จัดพิมพ์ เปรียบเทียบกับจำนวนเอกสารที่ตีพิมพ์ของผู้จัดพิมพ์นี้ และค่าที่นำมาใช้ต้อง Scale Normalized อยู่ระหว่าง 0 ถึง 1

4. วิธีการดำเนินการวิจัย

4.1 การเตรียมคลังเอกสารงานวิจัย

เอกสารงานวิจัยจาก academic.research.microsoft.com รวบรวมระหว่างเดือนมิถุนายนถึงสิงหาคม ปี 2013 ประกอบด้วยงานวิจัยจำนวน 71,828 บทความ โดยจำแนกเป็นบทความจากวารสารวิชาการ (Journal) 28,320 บทความ และบทความจากงานประชุมวิชาการ (Conference) 43,508 บทความ และในคลังเอกสาร (Document Corpus) ประกอบด้วยมีคำสำคัญ (Keyword) 23,073

งานวิจัยประกอบด้วย ชื่อ รายชื่อนักวิจัย บทคัดย่อ คำสำคัญ (Keyword) จำนวนเอกสารอ้างอิง (Reference) จำนวนผลงานถูกอ้างอิง (Citation) ปีและผู้จัดพิมพ์งานวิชาการ (Publisher) เช่น วารสารวิชาการ (Journal) 4,283 หรืองานประชุมวิชาการ (Conference) 2,885

4.2 การประเมินผล

วิธีการวัดประสิทธิภาพของ Indexing และ Ranking คือ ค่า Normalized Discounted Cumulative Gain (NDCG) โดย Jarvelin. Kekalainen [9]

การประเมินผลมาจากคะแนนผู้ใช้เป็นหลัก (Judgments) เรียกว่า Perfect Score นำมาคำนวณเป็น DCG Perfect คะแนนที่ได้ บ่งบอกว่าคำค้นมีความเกี่ยวข้องกับเอกสารนั้นๆ ที่ตำแหน่งที่ *k* เมื่อกำหนดให้คำค้น *q* และเซตเอกสารจากการค้นคืน คะแนนของเอกสารในแต่ละตำแหน่งสามารถคิดได้จากลำดับแรกจนถึงเอกสารลำดับสุดท้าย ตามสมการที่ (6)

$$NDCG_q = \sum_{j=1}^{k} \frac{(2^{r(j)}-1)}{log(1+j)} \quad (6)$$

เมื่อ *j* แทนตำแหน่งของเอกสาร และ *r(j)* แทนเลขจำนวนเต็ม ซึ่งเป็นค่าคะแนน (Judgment Score) ที่ได้จากผู้ทดสอบ

NDCG แทนค่าคะแนนความเกี่ยวข้องของเอกสารจากลำดับแรกสุดไปยังลำดับท้ายสุด

4.3 การทดสอบ

การทดสอบเพื่อพิสูจน์ตัวแบบที่นำเสนอ จึงจัดทำระบบ Research Paper Search Engine สำหรับเป็นหน้าเว็บของระบบสืบค้นที่เป็น Interface ให้กับผู้ทดสอบ

ในการทดสอบได้เชิญนักศึกษาระดับปริญญาโท ปริญญาเอก และนักวิจัยด้านวิทยาการคอมพิวเตอร์ กำหนดให้ผู้ทดสอบ

แต่ละคนใส่คำค้นที่ต้องการ เป็นคำ หรือประโยคใดๆ ก็ได้ในหน้าเว็บ ระบบจะสืบค้นข้อมูลจากทั้ง 4 ดัชนี โดยก่อนที่จะแสดงผลให้ผู้ทดสอบประเมิน ระบบจะมีการสุ่มลำดับและรวมผลลัพธ์ที่ได้เพื่อไม่ให้มีการแสดงเอกสารซ้ำในแต่ละดัชนี และเพื่อมิให้ผู้ทดสอบเกิดความลำเอียงในการให้คะแนนเอกสารที่ได้ในแต่ละลำดับด้วย

หน้าเว็บที่แสดงผลลัพธ์ โดยแสดงรายละเอียดบทความวิจัย ได้แก่ ชื่อ และบทคัดย่อ โดยผู้ทดสอบจะต้องอ่านรายละเอียดทั้งในส่วนของหัวข้อและบทคัดย่อ แล้วให้คะแนนบทความที่กำลังพิจารณาว่ามีความเกี่ยวข้องกับคำค้นมากน้อยแค่ไหน ซึ่งคะแนนที่ได้ดังกล่าวจะนำ ไปประเมินผลการเรียงลำดับที่ได้ (Research Paper Re-Ranking)

การทดสอบมีขั้นตอน ดังนี้

1. ผู้ทดสอบระบุคำค้นที่ต้องการในหน้าเว็บ

2. ระบบจะสืบค้นเอกสาร 30 ลำดับแรก ของแต่ละดัชนี โดยระบบจะตรวจสอบเอกสารที่แสดงผลซ้ำ และสุ่มลำดับการแสดงผลบนหน้าเว็บ เพื่อไม่ให้ผู้ทดสอบเกิดความลำเอียงในการให้คะแนนในแต่ละเอกสาร

3. ผู้ทดสอบให้คะแนนเอกสาร (Judgment Score) แต่ละเอกสารมีความเกี่ยวข้องกับคำค้นอย่างไร คะแนนอยู่ระหว่าง 4 ถึง 0 มีความหมาย ดังตารางที่ 1

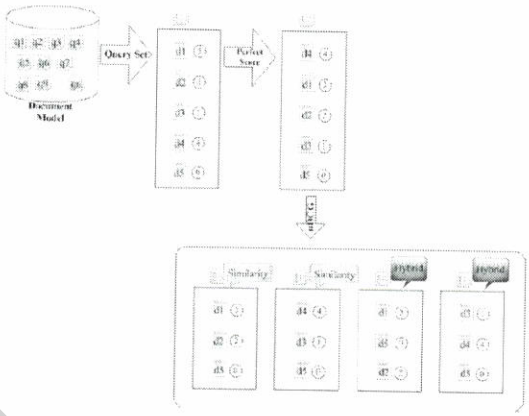
4. ระบบบันทึกข้อมูล

ตารางที่ 1 Judgment Score

คะแนน	รายละเอียด
4	มีความเกี่ยวข้อง (Extremely Relevant)
3	อาจจะเกี่ยวข้อง (Probably Relevant)
2	เกี่ยวข้องเพียงเล็กน้อย (Less Relevant)
1	อาจจะไม่มีเกี่ยวข้อง (Probably Not Relevant)
0	ไม่มีความเกี่ยวข้อง (Not Relevant)

5. ผลการทดลอง

จาก Judgment Score ที่ได้จากผู้ทดสอบ ในแต่ละครั้งของการสืบค้น คือข้อมูล 1 ชุดประเมิน แสดงดังภาพที่ 2 เอกสารทั้งหมดในหนึ่งชุดจะถูกเรียงลำดับตาม Judgment Score เพื่อหาค่า DCG Perfect Score และ ขั้นตอนที่ 2 เป็นการจำแนกเอกสารออกเป็น Index0 Index1 Hybrid0 และ Hybrid1 ได้ 4 ชุดดัชนี สำหรับหาค่า NDCG ที่ตำแหน่งการค้นคืนเอกสารที่ k



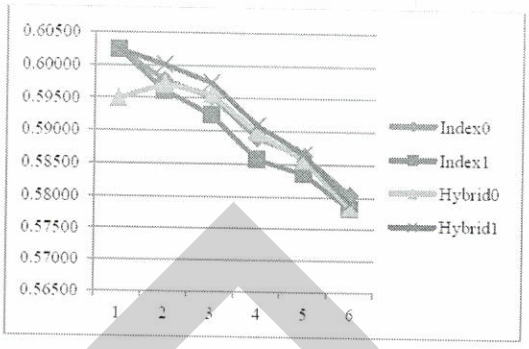
ภาพที่ 2: การคำนวณค่า NDCG

จากผลการทดสอบเบื้องต้น มีผู้ร่วมทดสอบ 20 คน โดยมีจำนวนคำค้นทั้งหมด 37 คำสืบค้น ในการประเมินผลหาค่าเฉลี่ย NDCG ในแต่ละดัชนี เมื่อแกน x แทนลำดับผลการค้นคืนใน 6 ลำดับแรก และแกน y แทนค่าเฉลี่ยของ NDCG ที่ได้จากการสืบค้น

เมื่อพิจารณาจากกราฟ แสดงดังภาพที่ 3 พบว่า NDCG ที่ตำแหน่งเอกสาร k=1 ที่ Index0 และ Hybrid1 ได้ค่า NDCG เท่ากับ 0.60238 เท่ากัน ดังตารางที่ 2 แต่เมื่อพิจารณาค่าเฉลี่ย NDCG ในช่วงตำแหน่งที่ k=2 ถึง 4 ค่าเฉลี่ย NDCG ของดัชนี Hybrid1 สามารถให้ผลการค้นคืนเอกสารที่ดีที่สุด โดยพิจารณาจากค่าเฉลี่ย NDCG ของผลลัพธ์

ค่าดัชนีของ Hybrid1 มีการกำหนด $\propto$ เท่ากับ 0.9 ตามสมการที่ (3) ซึ่งเป็นอัตราส่วนระหว่างค่าน้ำหนักของ Similarity กับ Static Ranking ที่มีส่วนช่วยให้อัตราการเรียงลำดับผลลัพธ์การค้นคืนดีขึ้น เมื่อเปรียบเทียบกับการเรียงลำดับบทความวิจัย ซึ่งส่วนใหญ่จะให้ความสำคัญกับความเหมือนมากกว่าคุณภาพของเอกสาร

ค่าน้ำหนักที่ได้จากการทดลอง เป็นผลของการนำเอา Static Feature มาช่วยในการเรียงลำดับ ส่งผลให้เกิดประสิทธิผลของการเรียงลำดับที่ดีขึ้นเมื่อนำมาใช้ร่วมกับ Similarity Feature เพียงอย่างเดียว



ภาพที่ 3: เปรียบเทียบค่า NDCG ของแต่ละดัชนี

ตารางที่ 2 ค่าเฉลี่ย NDCG

k	Index0	Index1	Hybrid0	Hybrid1
1	0.60238	0.60238	0.59497	0.60238
2	0.59776	0.59613	0.59715	0.60007
3	0.59524	0.59240	0.59561	0.59739
4	0.58898	0.58567	0.58980	0.59093
5	0.58630	0.58349	0.58565	0.58660
6	0.58042	0.57821	0.57848	0.57905

6. สรุปและอภิปรายผล

จากงานการทดสอบเบื้องต้นพบว่า เอกสารงานวิจัยที่มีจำนวนการอ้างอิงจากเอกสารอื่นมากกว่า เป็นงานที่มีการตีพิมพ์หรือเผยแพร่ออกมาแล้วในช่วงระยะเวลาหนึ่ง ซึ่งเป็นจุดที่ทำให้งานวิจัยใหม่หรืองานที่เพิ่งตีพิมพ์ออกมาปัจจุบันยังไม่สามารถถูกดันขึ้นมาแสดงผลการค้นคืนในลำดับต้นๆ ได้ ดังนั้นกรอบแนวความคิดนี้ จึงเหมาะสำหรับการสืบค้นเอกสารที่เน้นด้านคุณภาพมากกว่าความใหม่ของงานวิจัย เช่น ปีที่ตีพิมพ์ ซึ่งเป็นปัจจัยที่แปรผกผันกับจำนวนการถูกอ้างอิงจากเอกสารอื่นๆ

ค่า Judgment Score ที่ใช้เป็นพื้นฐานสำหรับคำนวณค่าที่บ่งบอกถึงความพึงพอใจของผู้ใช้งาน จำเป็นต้องเพิ่มปริมาณผู้ทดสอบเพื่อให้ได้ค่าเฉลี่ยที่ถูกต้องและผิดพลาดน้อยลง และสามารถแสดงผลค่าทางสถิติได้อย่างมีนัยสำคัญต่อไป

เอกสารอ้างอิง

[1] X.Geng, T. Y. Liu, and H. Li, "Feature selection for ranking," *Proc ACM SIGIR*, 2007.

[2] W. Choochaiwattana, M. B. Spring, "Applying social annotation to retrieve and re-rank web resources," *Internationalconference on information management and engineering (IEEE)*, 2009.

[3] W.Choochaiwattana "Usage of tagging for research paper recommendation," *3rd International conference on Advanced computer theory and engineering (IEEE)*, 2010.

[4] S.Noel, R.Beale "Sharing vocabularies: tag usage in CiteULike," *the British computer society*, 2008.

[5] D.H.Lee, P.Brusilovsky "Socail networks and interest similarity: the case of CiteULike," *ACM*, 2010.

[6] H.Wang, X.He, M.Chang, Y.Song, R.W.White, W.Chu "Personalized ranking model adaptation for web search" *ACM*, 2013

[7] D.Parra, P.Brusilovsky "Collaborative filtering for social tagging system: an experiment with CiteULike," *ACM*, 2009.

[8] M.S.Pera, Y.Ng "A personalized recommendation system on scholarly publications," *ACM*, 2011.

[9] P. Jomsri, "A combination of similarity ranking and time for social research paper searching," *World academy of science, engineering and technology 54*, 2011.

[10] P. Jomsri, S. Sanguansintukul, and W. Choochaiwattana "CiteRank: combination similarity and static ranking with research paper searching," *International Journal of Internet Technology and Secured Transactions*, Volume 3 Issue 2, April 2011 .

[11] K. Jarvelin, and J. Kekalainen, "IR evaluation methods for retrieving highly relevant documents," *Proc. ACM SIGIR conference on Research and Development on Information Retrieval*, July 2000.

[12] Q. Wu, C. J. C. Burges, K. M. Svore and J Gao, "Microsoft research technical report MSR-TR-2008-109," October 15, 2008.

ประวัติผู้เขียน

ชื่อ-นามสกุล	ขวัญเรือน โสอุบล
ประวัติการศึกษา	ปีการศึกษา 2544 สำเร็จการศึกษาระดับปริญาตรี สาขาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์ มหาวิทยาลัยบูรพา ปีการศึกษา 2547 สำเร็จการศึกษาระดับปริญญาโท สาขาวิทยาการสารสนเทศ คณะเทคโนโลยีสารสนเทศ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหาร ลาดกระบัง
ตำแหน่งและสถานที่ทำงานปัจจุบัน	IT Specialist บริษัท แอดวานซ์ อินโฟร์ เซอร์วิส จำกัด (มหาชน) ตั้งอยู่เลขที่ 1291/1 อาคารเอไอเอส ถนนพหลโยธิน แขวงสามเสนใน เขตพญาไท จังหวัดกรุงเทพมหานคร
ประสบการณ์ทำงาน	Technical Support Engineer Programmer System Analyst